A Comprehensive Survey on Rare Event Prediction

CHATHURANGI SHYALIKA, Artificial Intelligence Institute, University of South Carolina, Columbia, United States

RUWAN WICKRAMARACHCHI, Artificial Intelligence Institute, University of South Carolina, Columbia, United States

AMIT P. SHETH, Artificial Intelligence Institute, University of South Carolina, Columbia, United States

Rare event prediction involves identifying and forecasting events with a low probability using machine learning (ML) and data analysis. Due to the imbalanced data distributions, where the frequency of common events vastly outweighs that of rare events, it requires using specialized methods within each step of the ML pipeline, that is, from data processing to algorithms to evaluation protocols. Predicting the occurrences of rare events is important for real-world applications, such as Industry 4.0, and is an active research area in statistical and ML. This article comprehensively reviews the current approaches for rare event prediction along four dimensions: rare event data, data processing, algorithmic approaches, and evaluation approaches. Specifically, we consider 73 datasets from different modalities (i.e., numerical, image, text, and audio), four major categories of data processing, five major algorithmic groupings, and two broader evaluation approaches. This article aims to identify gaps in the current literature and highlight the challenges of predicting rare events. It also suggests potential research directions, which can help guide practitioners and researchers.

CCS Concepts: • **Computing methodologies → Artificial intelligence**;

Additional Key Words and Phrases: Event-prediction, rare-events, time-series, anomaly prediction, forecasting

ACM Reference Format:

Chathurangi Shyalika, Ruwan Wickramarachchi, and Amit P. Sheth. 2024. A Comprehensive Survey on Rare Event Prediction. *ACM Comput. Surv.* 57, 3, Article 70 (November 2024), 39 pages. https://doi.org/10.1145/3699955

1 Introduction

Events are incidents that are associated with specific locations (spatial), time periods (temporal), and contexts (semantics). Rare events are a subset of events that stand out due to their infrequency.

The degree of infrequency of rare events is typically influenced by the specific field of application [81, 92]. Rare event learning is considered an NP-hard problem [193], as it requires analyzing a large amount of data to identify rare events, which can be computationally intensive and time-consuming, especially in high-dimensional spaces. The size and complexity of rare event data lead to the challenge of handling this problem, resulting in complicated issues in data mining and ML. Imbalanced event datasets exhibit a prevalence of rare occurrences, wherein the quantity of instances associated with one class is significantly lower than the quantity of instances pertaining to the other. These datasets present challenges for learning algorithms since they may result in biased outcomes in downstream tasks such as classification, clustering, forecasting, and simulation. Algorithms necessitate tailoring to effectively address rare events, as these occurrences often give rise to challenges stemming from their uncommon nature.

Considering the significance of rare events across various domains, extensive research has been conducted in diverse knowledge areas, leading to a conflation of terms and issues within the literature. While terms like rare events, anomalies, novelties, and outliers may appear similar, it is crucial to note their distinct differences. Rare events and anomalies share characteristics such as an imbalanced class distribution and representation of all classes in the training set [56]. However, rare events primarily relate to temporal data, contrasting with anomalies that typically involve static data distributions. Novelties, on the other hand, involve the identification of new or unknown data patterns. Static novelties are characterized by supervised classification with a single class for training, whereas dynamic novelties entail supervised classification with an indeterminate number of labels [56, 68, 153, 174]. Outliers, distinguished from the aforementioned categories, are observations significantly deviating from the majority in the dataset [56, 126]. They often manifest in temporal data and are addressed through unsupervised classification methodologies. The primary focus of our investigation in this article is exclusively directed towards rare events. Hence, we specifically selected studies that include the term "rare events" in the article title and abstract.

In real-life, rare events can be observed ubiquitously in various domains, including medical diagnosis, fraud detection, and natural disaster prediction. In any field or area of study, rare events can be regarded as occurrences that possess valuable and meaningful information [147]. Rare events can be weighted by "*rarity*", a measure of being rare, uncommon, or scarce. Liu and Feng [124] formulate this as the "Curse of Rarity" (CoR). The fundamental idea behind CoR is that the events of interest are exceptionally rare, resulting in limited information in the available data. CoR leads to many underlying issues, including decision-making, modeling, verification, and validation. For instance, detecting rare diseases or medical conditions is challenging but crucial for effective diagnosis and treatment. Similarly, detecting fraud in financial transactions can help prevent financial losses and ensure the security of transactions. In the case of natural disaster prediction, identifying rare events such as earthquakes or tsunamis can help in effective disaster management and response. In manufacturing, these events lead to unplanned downtime or shutdowns, which are particularly detrimental for industries regarding equipment life and power consumption. Thus, exploring rare events in advance allows industries to implement mitigation procedures to reduce defects so that equipment downtime can be lowered, optimizing energy consumption and ensuring optimization, quality, and safety standards in processes.

The significance of rare events resides in their capacity to yield a disproportionate influence, surpassing that of more typical events. For instance, a disease that affects a minor percentage of the population can enormously affect public health. Similarly, large-scale fraudulent activities can result in substantial financial losses for individuals and organizations. Hence, it is evident that rare occurrences require specialized attention and analysis. Therefore, developing effective methodologies and algorithms that can handle the uniqueness and mitigate the biases and limitations inherent in rare events is essential. The problem of imbalanced datasets and rare events

is not new, and researchers have developed several techniques to address this issue. These techniques range from data-level approaches, such as oversampling and undersampling [40, 104, 160, 170, 198, 211], to algorithm-level approaches, such as cost-sensitive learning [113, 120, 145, 211] and ensemble methods [65]. In recent years, deep learning methods [41, 45, 51, 97, 133, 146, 159, 202] have also been applied to address the problem of rare events.

This survey article aims to provide a comparative review of the existing literature on rare event prediction. We have developed a taxonomy that summarizes the research on rare event prediction into four categories: rare event data, data processing, algorithmic approaches, and evaluation approaches. The rare event data section identifies datasets containing rare events, while the data processing section emphasizes the vital role of data processing in handling rare event datasets, refining data quality to enhance predictive model performance. The algorithmic approaches present mathematical models applicable to diverse scenarios encompassing various use cases, and the evaluation section delves into the multifaceted evaluation criteria utilized in rare event prediction studies. Then, we explore the existing literature pertaining to these four categories. The extended review involves the reviewing of each of these four categories with levels of rarity that we introduce, established industries, types of datasets, data modalities, applications, and downstream tasks. Finally, we highlight some open research questions and future directions identified in this area.

1.1 The Contributions of this Article

The main contributions of this article include:

(1) Comparatively review the existing literature on rare event prediction in four approaches; rare event data, data processing, algorithmic, and evaluation approaches.
(2) Analyze the literature by examining multiple avenues: dataset types, modalities, and downstream tasks.
(3) Identify gaps, challenges, and special concerns in the current research landscape while discussing potential emerging trends in the realm.

1.2 Organization

This study distinguishes a four-fold summarized categorization of approaches to learning from imbalanced event data for rare event prediction that have been implemented within related work. As shown in Figure 1, the main groupings include (I) Rare event data, (II) Data processing approaches, (III) Algorithm level techniques, and (IV) Evaluation approaches.
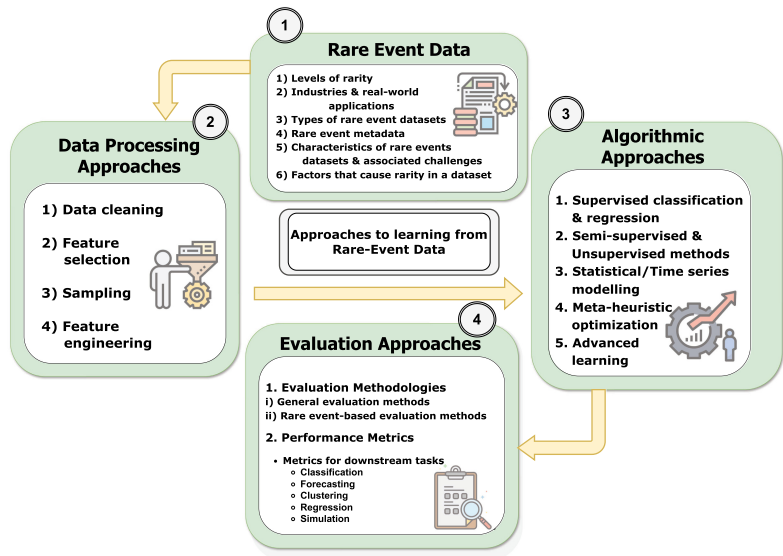

Fig. 1. Approaches to handling from rare event data.

2   Rare Event Data

This section analyzes datasets with rare events. Considering a wide range of datasets from multiple industries and different modalities we first categorize them by the rarity percentage. We then explore real-world applications of the data and examine data acquisition methods. Next, we present an analysis considering types of datasets, metadata, and modality. Finally, we discuss the characteristics and challenges of handling such data and the factors contributing to their rarity.

2.1   Datasets with Rare Events – Analysis of Existing Datasets with Rare Events

2.1.1   Levels of Rarity. In any domain, the rarity of events is inversely correlated with the maturity of that industry. At the same time, rarity is correlated with event frequency or the probability of occurrence. For better understanding and analysis, we introduce the notion of *"levels of rarity"*, which categorizes rarity into four levels as depicted in Figure 2 . These rarity levels will be used throughout all the review sections in this article. We have established the boundaries for the levels based on several factors, such as the distribution of data, significant differences in rarity levels, and context. The R1 category comprises datasets with events that have a frequency of 0–1%, which are considered extremely rare. On the other hand, events with a rarity of 1–5% are classified as very rare and fall under the R2 category. The R3 category includes events with a rarity of 5–10%, which are considered moderately rare. Finally, the rest of the events with a frequency greater than 10% belong to the frequently-rare (R4) category. It is apparent that when moving up in levels, the percentage of rarity and the event frequency tend to decrease, necessitating more sophisticated approaches for identifying and analyzing. The datasets in the studies we reviewed utilized diverse data types, including numeric (N), textual (TX), image (I), and audio (A). A special property of these "rare event" datasets is their adherence to the temporal nature (T) of "events" and/or adaptation of time-dependent features. We have encountered a limited number of datasets (2) that deviate from the typical time series data used in rare event studies, which we have excluded from our analysis. Consequently, we delineate the scope of rare events as occurrences within a time series, where an entire or part of a time series may constitute a rare event evolving over an extended duration.

2.1.2   Industries and Real-world Applications. The different industries that have been identified in this review include eight main sectors: economy, healthcare, transportation, telecommunications, manufacturing, energy, earth science, and others. It should be noted that the primary industries include several application domains that we merged into a general main industry, as shown in Table 1. In Table 1, we have compiled a list of real-world applications of rare events categorized by industry, application domain, and rarity based on the literature. Notably, these applications have predominantly concentrated on use cases such as detection, diagnosis, prediction, and downstream tasks like classification (CF), clustering (CL), forecasting (FT), regression (RG), and simulation (SM).

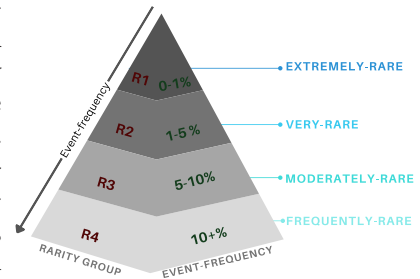Fig. 2. Levels of rarity.

2.1.3   Types of Rare Events Datasets. The study identified three types of rare event datasets: naturally rare event datasets, derived datasets, and simulated/synthetic datasets.

(i) Naturally rare event datasets (RE): These refer to datasets that inherently exhibit a low occurrence rate of specific events or phenomena. Table 3 (Appendix A.1) summarizes naturally

Table 1. List of Selected Rare Event Applications

Industry	Applications Domains	Examples from literature
Earth Sciences	Environment management, Disaster management, Geology	Detection of changes in buildings from aerial images taken before and after a tsunami disaster ([77, 89] – R1) O_3 state prediction, Prediction of hazardous seismic bumps in coal mines ([120] – R3) Detection of oil spills in satellite-borne radar images ([113, 132] – R2) Prediction of landslides, undesirable events in offshore oil wells ([29] – R4)
Manufacturing	Machinery fault diagnosis, Anomaly detection	Prediction of paper breaks in the paper manufacturing industry ([160, 161] – R1) Prediction of which parts will fail at the end of the production line ([96] – R1)
Telecommunication	Network monitoring, Telecommunication failure diagnosis	Prediction of telecommunications equipment failures ([33, 195] – R1) Prediction and classification of rare events in power grids ([6, 14] – R1, R2, R3, R4)
Transportation	Traffic monitoring, Autonomous driving, Aircraft health management	Degradation detection and trending, and failure discrimination based on the classification of aircraft systems ([41] – R1) Prediction of wrong-way driving in highways ([45] –R1)
Economy	Market trend forecasting, Fraud detection	Prediction of black-swan events in the Indian stock market ([51] –R1) Predicting credit card frauds ([145] – R1)
Healthcare	Disease diagnosis, Emergent incidents identification, Detecting microcalcifications	Post-Traumatic Seizure Detection ([162] – R1) Detection of LASA cases and rare events' classification in imbalanced healthcare data ([211] –R4) Detection of microcalcifications in mammography ([59] – R2)
Energy	Energy forecasting	Prediction of when to expect a high KWh cost ([49] –R3)
Others	Scene understanding, Computer vision, Remote sensing, Classification, Criminological investigation	Detection of scene changes from street view panorama images ([89] – R1) Detection of recidivists from criminological events ([98] – R4) Classification of types of glass in criminological events ([114, 191] – R4) Detection of a change of digit from a pair of samples in MNIST ([89] – R1, R4)

rare event datasets used in rare event literature. It was noticed that approximately 41% of the naturally rare datasets used fall into the extremely-rare category. Naturally rare event datasets can be sourced from various industries and domains, such as manufacturing, healthcare, earth sciences, and economy. For instance, data on environmental disasters such as oil spills [113] and tsunamis [77] can be gathered using satellite, ariel, or drone imagery via remote sensing technologies, while machine faults and anomalies can be tracked through data collection from sensors attached to different parts of the machines in manufacturing plants [23, 160]. Similarly, sources for predicting rare economic events may include financial market data, economic indicators, government reports, transaction data, consumer behavior data, social media data, and other relevant sources specific to the economic domain [51]. Data collected from such sources can be obtained from public databases. Repositories like **University of California Irvine (UCI)** [2], **Knowledge Extraction based on Evolutionary Learning** (KEEL) [7], **Outlier Detection DataSets (ODDS)** [164], Kaggle [12], research-specific data storages [11, 14, 20, 23, 29, 80], industry-specific databases [1, 10, 26, 27, 30, 33, 35, 141, 142], **Application Programming Interfaces (APIs)** [8, 31, 75], social media platforms [34, 99], and news outlets [34, 99] would be such databases. Data extraction from them can involve manually accessing public datasets, querying databases, web scraping, or using web services.

(ii) Derived datasets (DE): A derived dataset is a dataset that is not initially rare, results from transformation of an existing dataset, and includes a new insight, "derived rarity". For this, the original dataset should ideally possess sufficient information and features related to multiple events of interest and should initially be not rare. The availability of such data enables performing operations, calculations, or algorithms on the raw dataset to transform it into a derived dataset that captures the "derived rarity" aspect [41, 67, 120]. One example of derived datasets is the **ABCD** (**AIST Building Change Detection**) dataset [15, 16, 77], which was initially not designed to include rare events but has been manipulated in certain studies to create rare event scenarios

by altering event percentages [40, 60, 89]. MNIST dataset [85], originally an image dataset, has been repurposed in studies to detect letter changes, creating an imbalanced dataset by reducing the frequency of certain letters [40]. Table 4 (Appendix A.2) summarizes derived datasets used in rare event research in the studied literature. The analysis findings indicate that a substantial portion (34%) of the derived datasets we considered in this review belong to the extremely rare category.

(iii) Simulated/Synthetic (SIY) datasets: A simulated (also termed synthetic) dataset is an artificially generated dataset that mimics the characteristics and patterns of real-world rare events. They are typically created based on known models, distributions, or algorithms to replicate the statistical properties and relationships observed in the original data. These data can be generated in cases where collecting direct data on rare events is challenging and impractical. They are beneficial for predicting rare events that have not yet occurred or for testing the accuracy of predictive models in a controlled environment [29, 60, 114, 132, 146]. In the literature, SIY datasets have been generated in controlled environments using tools such as MATLAB, OLGA Dynamic Multiphase Flow Simulator [32], **Signal Fragment Assembler (SFA)**, **Variational Autoencoder (VAE)**, **Data Picker (DP)**, **Quality Classifier (QC)** [57, 58, 107], and frameworks like Massive Online Analysis (MOA) [52, 120], as presented in Table 5 (Appendix A.3). While most studies focus on naturally rare and derived datasets, considerably less research is based on simulated/synthetic datasets.

2.1.4 Rare Event Metadata. Metadata refers to the descriptive information that provides additional context about a dataset. It includes information about the structure, format, quality, source, and other data characteristics. In rare event data acquisition, metadata acquisition involves extracting relevant information from the data that contributes to a better understanding of the rare events. While metadata may be applicable to both rare and majority classes, its importance for rare events stems from the need to capture additional attributes specific to these events. This can include attributes such as timestamps, geographical location, variables related to the event's occurrence, data sources, data collection methods, and any other contextual information that aids in analyzing and interpreting the rare events. The following methods were revealed as ways of acquiring metadata.

   (i) Data collection: Metadata can be collected during data acquisition. This includes capturing information such as the source of the data, timestamps, geographical coordinates, sensor settings, or any other relevant contextual details that help describe the data [48, 51, 67, 109].
  (ii) Expert opinions: Engaging domain experts or subject matter specialists can provide valuable metadata. Experts can contribute their knowledge about rare events, their causes-effects, associated variables of interest, equations, hypotheses, or factors that influence their occurrence [29, 60, 67, 132, 146, 187]. This knowledge can assist in identifying appropriate metadata to enhance the analysis and prediction of rare events.
 (iii) Data annotation: Adding annotations or labels to the data can serve as metadata. This requires manually categorizing data instances as rare events or non-rare events, assigning event severity or damage levels, utilizing standard metrics and indexes, and providing additional descriptive labels that capture specific attributes or characteristics of the rare events [38, 49, 51, 146, 187, 206].
 (iv) External sources: Documents, technical reports, publications, and websites offer insights, statistical data, or contextual details that contribute to the understanding and analysis of rare events. For example, clinical reports, insurance claims data, state mortality records from government websites, outpatient visit details from electronic health records, and

responses to patient-reported measures like the **Patient Health Questionnaire** (**PHQ-9**) have provided valuable information in predicting rare medical incidents [63, 176].

2.1.5 *Characteristics of Rare Event Datasets and Associated Challenges.* Rare event datasets exhibit distinct characteristics that lead to various related issues, some outlined below.

(i) Skewed class distribution and lack of data: A skewed class distribution is a distribution of classes that is not symmetrical or evenly distributed. Class imbalance is a specific case of skewed class distribution featuring a substantial disparity in the number of instances between classes [137, 192]. In rare events datasets, the minority class has a significantly smaller number of samples than the majority class. This skewness in class distribution makes it difficult for ML algorithms to learn patterns and classify the minority class accurately [113]. Lack of data can take two forms. Absolute rarity occurs where the number of samples associated with the minority class is small in the absolute sense, whereas relative rarity happens where the minority class samples are less relative to the other classes [194]. These rarity forms pose consequential challenges for classifiers in identifying patterns and regularities within these rare occurrences to learn a robust model [113, 146].

(ii) Temporal property: An inherent characteristic of rare event datasets is the temporal aspect, which refers to the occurrence or sequencing of events over time. Temporal property is essential in analyzing and understanding rare event data, as it provides insights into the timing, duration, order, and interdependencies of rare events and normal events [56]. However, due to the imbalanced class distribution and data sparsity in a temporal context, capturing the time-dependent patterns and correlations in data can be a major challenge. Thus, handling these issues in a temporal context adds complexity to the analysis, necessitating specialized techniques for accurate prediction.

(iii) Class overlap: In some cases, there may be overlapping patterns or similarities between the rare event(minority class) and the non-rare events(majority class) [140, 177]. This can lead to misclassifying rare events as more common, resulting in false negatives.

(iv) Uncertainty: Uncertainty, in the context of rare event datasets, refers to the lack of precise knowledge and confidence in the observed data. This arises from the limited sample size, data sparsity, high-class imbalance, and lack of information, and it becomes a challenging issue in generalizing the findings of any downstream task [51, 70].

(v) High dimensionality: Rare event datasets can include many features or dimensions, including numerical, categorical, or textual variables [82, 157]. Including multiple features complicates the modeling process and requires careful feature selection or dimensionality reduction techniques.

(vi) Event complexity: Rare events involve intricate relationships between multiple variables, which generate complex patterns and interactions in real-world systems [109]. This often makes modeling rare events difficult, necessitating sophisticated modeling approaches to capture the underlying complexity.

2.1.6 *Factors that Cause Rarity in a Dataset.* Several factors can cause rarity in a dataset, including:

(i) Natural occurrences: Some events simply occur less frequently in nature than others. For example, diseases that affect a small percentage of the population will naturally result in highly imbalanced datasets. Natural hazards like landslides, tsunamis, and seismic bumps or climate catastrophes like heavy flooding that may occur once in a century are also rare occurrences [120, 186, 206, 210]. Black-swan events, unpredictable events with severe global economic consequences, are extremely rare natural events in the economic sector.

Events like the 9/11 attack and the Chinese economic downturn are examples of these [51].

(ii) Class definition: The definition of a class can also impact its rarity. If a class is defined narrowly or specifically, it may contain fewer instances than if defined broadly [140]. For instance, when estimating the likelihood of stroke based on varied input parameters like demographic data, additional illnesses, and smoking status, the definitions of "having a stroke" and "not having a stroke" may be narrowly defined. Due to these specific criteria used in classification, stroke events within the dataset may be deemed rare.

(iii) Sampling bias: If data are collected from a biased sample, they may not precisely reflect the true distribution of the population. In particular, if a rare event dataset only contains events occurring in certain geographic regions or in certain types of people, this can cause rarity. For instance, while **HIV (Human Immunodeficiency Virus)** is not classified as rare due to its global prevalence [148], there are specific regions and subpopulations where HIV is relatively rare, indicating variations in its incidence and prevalence.

(iv) Cost and measurement errors: Data collection and labeling can be expensive, and in some cases, collecting large amounts of data for a rare event may not be feasible. Evidently, the generation of simulated data and synthetic data may be costly. It may be challenging to accurately observe and measure rare events in natural and controlled environments, resulting in fewer labeled instances. Similarly, rare events that involve complex interactions among their dimensions may not always be observable and measurable accurately. For example, rare events like **Wrong-Way Driving (WWD)** are influenced by multiple variables like roadway geometry and configurations, traffic volume, lighting conditions, weather conditions, and driver's age and medical conditions. However, the impact of these complex interactions is not always readily observable and measurable, giving rise to various complications [45].

(v) Subjective decisions: Curators can selectively include instances of a rare event based on specific criteria by selecting features and classes that impact the rarity of that class within the dataset. The decision to include only such instances can affect the rarity of the class within the dataset. This is apparent when creating new datasets by generating derived, simulated, or synthetic data based on the original rare event datasets. For example, in predicting rare events like extreme heat waves, curators have used subjective decisions by selecting temperature and humidity thresholds and time periods based on their understanding of heat wave severity [49]. These decisions can be made to categorize the data into classes and identify patterns related to moderate, severe, or extreme heat waves.

The relationship between rare event data, acquisition methods, rarity factors, characteristics, and challenges of rare event datasets is multifaceted and interconnected, as summarized in Figure 3. The factors contributing to a dataset's rarity inherently result in rare event data and their characteristics. Rare event data constitute rare event datasets, which can be acquired through various methods. The characteristics of these data, in turn, give rise to various challenges in analyzing and predicting rare events, as discussed in the previous section.

In conclusion, this section examined four significant subsections in depth: rare event datasets, data acquisition methods, factors of rarity, characteristics, and challenges of dealing with rare event data. We devised a rarity hierarchy that provided a systematic method for summing data pertinent to rare events for the purposes of analysis. While most datasets and studies fall under the extremely-rare category in the hierarchy of rarity, many research projects are based on naturally rare and derived datasets. Textual and audio data-based research on rare events has received less attention than time series, image, and numerical data-based datasets. It is worth revealing that rare
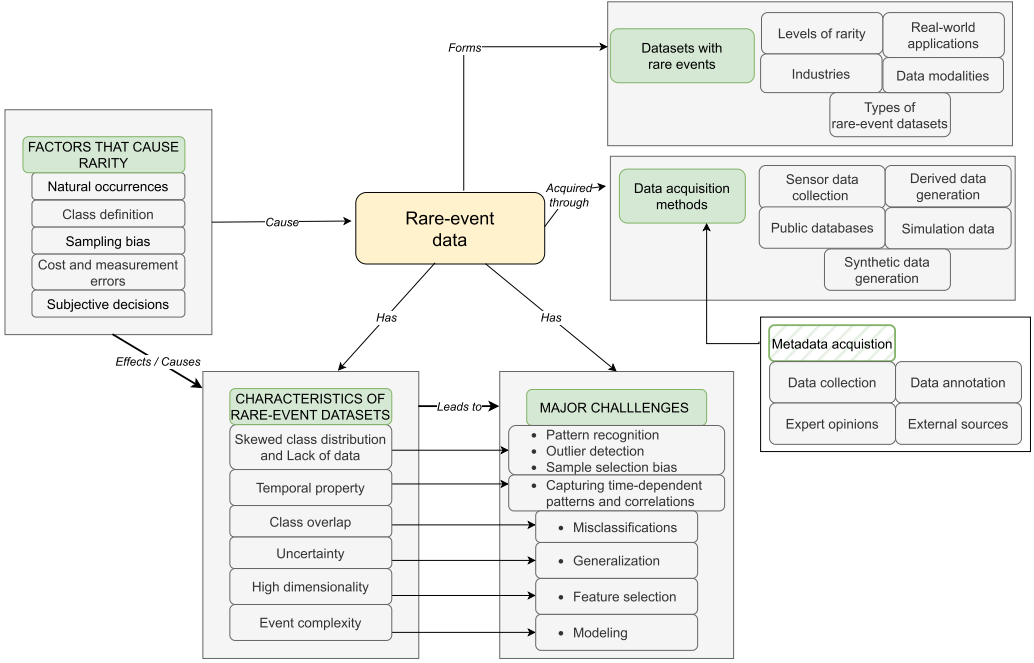
Fig. 3. Relationship between rare event data, acquisition methods, rarity factors, characteristics, and challenges of rare event datasets.

event-related problems and research are not restricted to a specific domain, industry, or sector, as we investigated research efforts across multiple sectors. Finally, the relationship between rare event data, acquisition methods, rarity factors, characteristics, and challenges of rare event datasets are drawn to summarize the overall scope of review in this section.

3 Data Processing Approaches

This section focuses on the importance of data processing methods in enhancing data quality for improved predictive model performance when dealing with rare event datasets. It explores various data processing approaches used in rare event prediction research, emphasizing their specific objectives. The subsequent discussion provides a detailed examination of each objective, including an analysis of how these approaches intersect with data modality, rarity groups, and downstream tasks.

3.1 Objectives of Data Processing Approaches

In the literature, we identified that data processing approaches aim to achieve four main objectives. Firstly, it is responsible for data cleaning, which improves the quality, consistency, and reliability of data analysis results. Secondly, it caters to feature selection by selecting the optimal variables by limiting the input variables to the model and utilizing only relevant features. Thirdly, it aids in sampling by modifying the data samples to balance the distribution and/or eliminating undesirable samples at the data level. Finally, data processing methods are applied in feature engineering to transform raw series data into a stable format suitable for modeling. In Figure 4, we classify the data processing approaches in rare event research that adhere to the aforementioned four objectives into four main categories. The approaches to data processing in rare event studies are summarized

Data Processing Approaches			
Data Cleaning	**Feature Selection**	**Sampling**	**Feature Engineering**

Data sifting
Data filtering
Imputation
Noise removal
Data framing
Text-based cleaning techniques
Image processing techniques
	Unsupervised methods
	  Correlation-based feature selection
	
	Supervised methods
	  Wrapper-based methods
	  Filter-based methods
	    Statistical measures
	    Feature importance methods
	  Intrinsic-based methods
		Basic sampling
		  Random oversampling
		  Random undersampling
		
		Advanced sampling
		  SMOTE Timeseries subsampling
		  ADASYN. Stratified sampling
		  SMUTE Audio data sampling
		  ENN Uncertainty sampling
		  NCL Choice-based or
		  NearMiss endogenous sampling
		  OSS
		  CBO
			Data augmentation
			Data discretization
			Encoding
			Feature Scaling
			Dimensionality reduction
			Fourier transform

Fig. 4. Data processing approaches in rare event research.

in Table 6 (Appendix B.5), along with their primary categories, rarity groups, data modalities, and subsequent tasks.

3.2 Data Cleaning (DC)

Data cleaning focuses on the important task of preparing and refining the data. This involves various techniques and processes to improve the dataset's quality, consistency, and reliability. This subsection examines various data cleaning approaches and discusses the application of them in various data modalities, rarity levels, and downstream tasks (See Figure 7 and discussion in Appendix B.1).

3.2.1 Approaches to Data Cleaning. Various approaches address data cleaning tasks, often depending on the specific data modalities. Some common data cleaning methods used in rare event-based data mining include data sifting, filtering, imputation, noise removal, audio-based preprocessing, text-based preprocessing, and image processing techniques.

(i) Data sifting: Refers to refining large volumes of data to identify the most relevant and important information. In rare event literature, sifting has been used to systematically sort through the data to identify and extract specific subsets of data. There are two types of data sifting: heuristic data sifting and statistical data sifting. Heuristic data sifting relies on expert knowledge and intuition, while statistical data sifting relies on quantitative measures and algorithms to identify relevant features [60, 88, 117, 134]. In [60], important knowledge about Ozone (O_3) concentration (i.e., time periods where O_3 is high) has been used as a heuristic, and several statistical measures were undertaken to identify relationships between different O_3 states and related components.

(ii) Data filtering: Involves segregating and removing unwanted, irrelevant data or information from a large dataset. This includes techniques like removing duplicate data [97], removing records containing very small number of records per group [201, 202], removing rows containing irrelevant types [134], and removing rows that meet specific conditions [200]. For instance, in [134], only the connected calls were used, discarding the unconnected calls from VoIP traffic data in unsupervised rare event detection in spatiotemporal environments. [200] used a cut-off time to define an event of interest and excluded subjects censored before the cut-off time.

(iii) Imputation: Imputation or value approximation techniques estimate and represent missing and incorrect values in a dataset with reasonable approximations based on the available data. These techniques are vital for addressing the sparse, incomplete, and imbalanced nature of the rare event data, requiring careful consideration to avoid skewed results. Simple methods like mean and median imputation, though commonly used, can introduce bias in rare event scenarios due to the skewed nature of the data [17, 20, 73, 82, 157, 200]. Advanced techniques such as Interpolation techniques [40, 156], Iterative Imputation [38, 147, 178], **Multiple Imputation by Chained Equation (MICE)**[157], Soft Impute [207], **Expectation maximization (EM)** [71, 147], and **Singular Value Decomposition (SVD)** [184, 193] are better suited for rare-event datasets, as they account

for complex patterns and correlations [38, 193, 207]. These methods help maintain the integrity of rare events during the imputation process, making them significant for accurate predictions.

(iv) Noise removal: Involves eliminating random and irrelevant data variations that do not contribute meaningful information to patterns and relationships. The Brooks-Iyengar algorithm is a fault-tolerant method for sensor fusion, handling faulty sensor readings effectively [54]. Iyer et al. [100] utilized Brooks-Iyengar along with ensemble data cleaning trees to handle random noise and missing data. Sampling methods like **Tomek-links** (**TL**) and **Edited Nearest Neighbors** (**ENN**) are used for noise removal. TL removes noise and boundary points in the majority class during rare events [45, 111, 113, 120], while ENN eliminates noisy data samples, resulting in smoother decision boundaries [45].

(v) Text-based cleaning techniques: Standard text preprocessing techniques, such as converting text to lowercase, removing stop-words, and employing stemming and lemmatization, have been used in rare event prediction studies involving textual data. This standardization improves the reliability of predictions by addressing textual inconsistencies, reducing noise and data variability in rare events involving textual data [197, 211]. For instance, in healthcare, R packages such as "tm", "snowball", and "Rstem" have been used to standardize text data by normalizing word variations through stemming and lemmatization, thereby enhancing the prediction of rare medical incidents [197].

(vi) Image processing techniques: In rare event prediction using images, image processing techniques have been employed to clean and normalize the images. These techniques aim to detect suspicious regions and extract relevant features that differentiate rare regions from similar ones. Kubat et al. [113] used various image processing methods to correct for radar beam incidence angle, identify dark regions, and extract specific features like the size of oil spills and average brightness. The output of the image processing is a fixed-length feature vector for each suspicious region, facilitating further analysis and prediction.

3.3 Feature Selection (FS)

Feature selection aims to identify a subset of input features from a dataset to extract the most pertinent information and improve the model's predictive capacities by reducing complexity. This subsection examines various approaches for feature selection in rare event studies and discusses the application of these techniques to various data modalities, rarity levels, and downstream tasks (See Figure 8 and discussion in Appendix B.2).

3.3.1 Approaches to Feature Selection. Feature selection in rare event prediction is crucial due to the imbalance and sparsity of the data, requiring careful adaptation to identify the most relevant features while preserving critical information related to rare events. The techniques can be categorized into unsupervised and supervised methods.

(I) Unsupervised methods: Unsupervised feature selection methods are applied without considering the response or target variable, focusing on the relationships and patterns within the independent variables. For example, Correlation-based feature selection, frequently used in rare-event studies, evaluates the relationships between features using correlation coefficients like Pearson correlation, **Analysis of Variance** (**ANOVA**), and chi-squared tests. However, given the rarity of the events, thresholds for eliminating features must be set carefully to preserve variables that may appear redundant in general datasets but are critical in rare-event contexts [67, 120, 136, 146].

(II) Supervised methods: These methods use the response/target variable in the feature selection process and eliminate irrelevant variables in making a prediction. The supervised methods can be categorized into wrapper-based, filter-based, and intrinsic-based. Wrapper-based methods search for well-performing subsets of input features. They wrap or embed an ML algorithm within

their core to perform feature selection. In rare-event scenarios, wrapper-based methods like **Recursive Feature Elimination** (**RFE**) can be adapted to preserve features critical for identifying rare events, even if these features do not show strong significance in a broader dataset. For instance, RFE has been combined with **Hidden Markov Models** (**HMMs**) to predict rare events, as it is sensitive to the subtle temporal patterns that may indicate these uncommon occurrences [165]. Standard filter-based methods extract or select features based on their statistical relationship with the target variable. Given their rarity, traditional statistical measures can be misleading or insufficient to identify rare events. For instance, sliding window techniques and wavelet transforms can be adjusted to focus on detecting rare spikes or patterns that could be easily overshadowed by more frequent but less relevant data [203]. The adaptation of methods like wavelet analysis [203], **Discrete Wavelet Transform** (**DWT**) [165] and Gumbel copula function [206], **Minimum Redundancy Maximum Relevance** (**MRMR**) [136], **Term Frequency-Inverse Document Frequency** (**TF-IDF**) [197, 211] as seen in rare-event studies involves fine-tuning these techniques to capture the unique, low-frequency signals that often characterize rare events. Furthermore, statistical methods ranging from simple calculations like mean, median, variance, skewness, kurtosis, and standard deviation to more advanced measures like spectral energy and frequency entropy derived from time and frequency domain analyses have been used in rare event prediction in the financial domain [51, 144]. In rare-event prediction, feature importance methods require special attention to ensure that features relevant to rare occurrences are not overlooked. Methods like Gini importance [82, 96, 147] or XGBoost's accuracy and cover measurements [96] have been adapted to emphasize features that, while they might contribute minimally to overall model accuracy, are critical in predicting rare events. For example, the application of **Mel-frequency Cepstral Coefficients** (**MFCC**) in rare-event contexts goes beyond traditional audio signal processing, aiming to capture subtle acoustic variations that could signify the occurrence of an event, as demonstrated in studies focusing on audio-based rare-event prediction [36, 173]. Intrinsic methods, such as attention mechanisms, are particularly valuable in rare-event prediction due to their ability to dynamically focus on the most relevant features across different scales and dimensions. For instance, multi-variate and multi-scale attention methods have been specifically adapted to enhance the prediction of rare events by focusing on spatial-temporal features that are often subtle and dispersed [162]. Additionally, decision trees, which are intrinsically capable of feature selection, have been adapted to handle the unique distributional characteristics of rare-event datasets, ensuring that the rare but critical branches of the tree are not pruned away during model training [113, 120].

3.4 Sampling (SL)

Sampling techniques in ML are essential methods for selecting a subset of data from larger datasets and are often used to tackle challenges like class imbalance and large data sizes while enhancing computational efficiency. The application of rare event data showcases their effectiveness in significantly improving performance in mining rare events, highlighting their importance in research contributions. This subsection examines various approaches for sampling and discusses their application to various data modalities, rarity levels, and downstream tasks (See Figure 9 and discussion in Appendix B.3)).

3.4.1 Approaches to Sampling. Sampling techniques used in rare event prediction can be categorized into basic and advanced methods based on the complexity and sophistication of the sampling approaches utilized.

(I) Basic sampling techniques: These techniques seek to address the issue of class imbalance by eliminating instances from the majority class or increasing the minority class by duplicating

minority class samples. **Random minority oversampling (ROS)** and **random majority under-sampling (RUS)** are the most frequent basic sampling strategies. In ROS, instances of the minority class are replicated randomly in the dataset, while in RUS, occurrences of the majority class are randomly eliminated from the dataset. In rare event-based research, many studies have used ROS and RUS methods to achieve a more balanced class distribution [40, 104, 160, 170, 198, 211]. Some researchers have combined sampling techniques with clustering models [145, 198] and ensemble learning methods [59] to improve predictive performance. Some others have applied randomly over/under-sampling in conjunction with advanced architectures, such as the Siamese CNN [89], to enhance rare event detection. Additionally, statistical sampling methods, like Hoeffding bounds, have been employed in rare event learning using associative rules and higher-order statistics [100]. Even though basic sampling has been extensively used, it has several limitations. The drawback of the oversampling method is that it leads to overfitting since the model learns from the same duplicated samples repeatedly. Undersampling eliminates lots of data that could have been utilized to train the model and improve its accuracy.

(II) Advanced sampling techniques: Going beyond basic random adjustments, advanced sampling techniques utilize intelligent mechanisms that consider the distribution of data points and the nuance of learning specific examples, resulting in greater effectiveness for handling complex, imbalanced datasets.

(1) Synthetic minority oversampling technique (SMOTE): SMOTE is a popular advanced sampling technique used in ML to address class imbalances. It generates new synthetic minority cases by extrapolating from existing minority instances, and it considers the difference between a sample and its closest neighbor to create synthetic examples. SMOTE has been adopted in various rare event-related use cases implementing different modeling techniques [42, 45, 73, 121, 175, 211]. It has been observed that logistic regression combined with SMOTE produces good results in detecting **Look-Alike-Sound-Alike (LASA)** cases in textual data [211], but it can be computationally expensive. Adaptive swarm balancing algorithms [121] and dynamic churn prediction frameworks [42] have also utilized SMOTE for rare event prediction in imbalanced datasets. Additionally, Borderline-SMOTE is a variation that selectively applies SMOTE to minority instances on the border of the minority decision region, yielding effective results in mining tasks [170]. While SMOTE has been adopted in various rare event use cases, it has a significant drawback due to the arbitrary generation of synthetic data. As a result, the class boundaries between the majority class and the minority class in the synthetic data may appear significantly different from those in the original dataset, potentially deviating from the actual distribution of the minority class.

(2) Adaptive synthetic sampling (ADASYN): ADASYN [94] is a method that solves the issue with SMOTE by following a weighted distribution for minority classes according to their level of difficulty in learning. It generates synthetic observations of the harder-to-learn minority samples compared to the easier-to-learn minority samples and adaptively shifts the decision boundary towards the harder-to-learn samples. Asraf et al. [45] have used ADASYN with the XGBoost model in rare event modeling in highrisk WWD roadway segment identification.

(3) Similarity majority undersampling (SMUTE): SMUTE [120] is an undersampling technique that distinguishes between the majority and minority class samples by considering the cosine similarity between each and its neighboring minority class samples. SMUTE works by calculating similarity scores between each majority class sample and a given number of minority class samples, then selecting a subset of majority class samples with the highest percentage of high similarity scores based on a specified undersampling rate.

(4) Edited nearest neighbor (ENN): ENN [183] is an undersampling method that reduces noise and refines decision boundaries in imbalanced datasets. ENN eliminates samples whose class labels differ from most of their nearest neighbors, resulting in a dataset with smoother decision

boundaries and reduced noise. It has been applied with SMOTE oversampling to create a balanced, noise-free training dataset for improved rare event prediction [45, 120].

(5) Neighborhood cleaning rule (NCL): NCL is an undersampling technique that removes redundant, noisy, or ambiguous samples. NCL employs the ENN technique to remove the data samples [47]. However, compared with ENN, NCL has the additional benefit of removing redundant instances based on feature space similarity [120].

(6) NearMiss (NM): NM [130] is a technique that uses distance measures of majority class samples to minority class samples in selecting samples. When two points from different classes are located very close to one another in a distribution, this algorithm eliminates the data point from the larger class to balance the distribution. NearMiss-2(NM2), a variation of NM, has been used in rare event prediction data processing [45, 120].

(7) One-sided selection (OSS): OSS is an undersampling technique that balances imbalanced datasets by removing redundant and noisy majority-class instances. It combines the use of Tomek Links to identify ambiguous points on the class boundary with the **Condensed Nearest Neighbor** (**CNN**) rule [93] to eliminate distant redundant examples from the majority class, resulting in a minimally consistent subset without compromising model performance. OSS has the advantage over other methods of intelligently locating and removing redundant and noisy majority class instances by utilizing Tomek Links and the CNN rule. This makes a more refined and representative subset of the majority class. OSS has been employed in various rare event studies to address the high imbalance in learning tasks [59, 113, 114, 170].

(8) Cluster-based oversampling (CBO): CBO employs resampling by clustering each class's training data separately and then performing random oversampling, cluster by cluster [104, 170]. CBO employs separate clustering of each class's training data followed by random oversampling within each cluster, enhancing the representation of rare classes in a balanced manner. The advantage of this method is that it considers both between-class and within-class imbalances and then oversamples the data to correct both imbalances simultaneously.

(9) Time series subsampling: This involves selecting a subset of time points from the original to create a new, shorter, and more balanced time series [64, 78]. The difference between subsampling and time series subsampling is that subsampling is a general term that refers to selecting a random subset of data points, while time series subsampling is a specific type of subsampling used for time series data involving the selection of a subset of time points from a time series. In industrial-based research on machinery fault diagnosis [123], subsampling of time series has been performed to consider the balanced time series length of normal and different bearing fault types. Further, time series subsampling reduces the computational cost of analyzing long time series and aids in training ML models.

(10) Stratified sampling: It is a general sampling technique that divides the dataset into subgroups (strata) based on the target variable's classes and then randomly samples from each stratum to ensure representation of all classes. To sample normal and different bearing fault failure types, [123] uses stratified techniques to represent these subgroups in the final dataset properly.

(11) Audio data sampling: Data framing has been used as a data sampling technique for audio data [36] that involves converting the audio data into a machine-readable format to fix the audio file sampling (frame) rate. In this method, the audio data is divided into frames, each representing a segment of the audio signal sampled at a specific rate. The total number of frames can be calculated by multiplying the sampling rate by the audio file's duration.

(12) Uncertainty sampling: It is a data sampling technique widely used in active learning, particularly relevant to rare event prediction research. It entails selecting instances from a dataset based on the uncertainty or low confidence of their predicted labels by a ML model. The goal is to prioritize sampling data points for which the model lacks certainty in its predictions. This approach

is valuable in scenarios where labeling data is resource-intensive, such as in rare event prediction [152], as it allows researchers to actively select the most informative instances for annotation, thereby improving the model's performance with limited labeled data.

(13) Choice-based or endogenous sampling: Endogenous sampling is a rare event prediction method that selects samples based on the dependent variable (y) rather than the independent variable (X). It aims to obtain a representative sample that accurately reflects the distribution of rare events in the dataset, addressing the imbalance issue and improving predictive performance. Choice-based or endogenous stratified sampling has been used in various applications, such as **Light Detection and Ranging** (**LIDAR**) maps, where non-landslide cells are sampled one to five times more than landslide cells to achieve better representation [186]. Variations of regression models have been experimented with in endogenous sampling approaches [127].

3.5 Feature Engineering (FE)

Feature engineering involves converting raw data into a relevant feature set for modeling. It aims to extract meaningful information from the dataset and present it in a format suitable for the learning algorithm. This subsection examines various approaches to feature engineering in rare-event contexts. Then, it discusses their application to various data modalities, rarity levels, and downstream tasks (See Figure 10 and discussion in Appendix B.4).

3.5.1 Approaches to Feature Engineering. Investigating rare events, we will review commonly utilized techniques in feature engineering as outlined below.

(I) Data augmentation: In rare event prediction, data augmentation has been important in addressing the scarcity of event samples and the challenges of imbalanced datasets. Unlike general applications, where augmentation primarily boosts generalization, in rare events, it must be tailored to enhance model sensitivity to infrequent patterns. For instance, in the pulp-and-paper industry, a tailored augmentation method combined with **Fast Fourier Transform** (**FFT**) has been used to enrich time series data with synthetic rare event examples, improving fault detection [160, 161]. **Wasserstein GANs** (**WGANs**) have been adapted for rare event prediction by generating high-quality synthetic data to address data scarcity and improve training stability, while **Conditional GANs** (**CGANs**) are used in creating targeted samples based on specific attributes or conditions [73]. Apart from those, standard image-based augmentation techniques, like cropping, have been applied to create additional training samples while preserving the critical spatial features necessary for detecting rare mineral deposits, predicting mineral prospectivity, and identifying scene changes [89, 150, 205]. These targeted approaches ensure that models remain sensitive to rare events, improving their prediction accuracy.

(II) Data discretization: Data discretization can manage the unique challenges of skewed class distributions and high-dimensional data. Converting continuous data into discrete categories can help highlight the rare event patterns that might otherwise be overshadowed in continuous data [147], simplifying the model's ability to detect and learn from rare occurrences. Standard techniques like optimal binning have allowed supervised discretization that preserves essential characteristics of rare event data, ensuring that the model captures the subtle distinctions necessary for accurate prediction [49, 144].

(III) Feature scaling: In rare event research, standardization has been used to align feature distributions with a standard normal distribution, which helps in mitigating the effect of class imbalance by ensuring that features do not disproportionately influence the model based on their scale [36, 119, 123, 202]. Normalization rescales values to a uniform range, preventing features with larger ranges from dominating the model's learning process [132]. Both techniques ensure the model's sensitivity to minority class features.

(IV) Dimensionality reduction: Dimensionality reduction is more than a standard preprocessing step; it is essential for rare event prediction as it addresses the high dimensionality and complexity inherent in rare event datasets. **Principal Component Analysis (PCA)** reduces the feature space by preserving the most informative components that distinguish rare events from normal occurrences [36, 41, 51, 73, 132, 147, 162, 165, 202, 203]. Focusing on the principal components, PCA filters out noise and highlights subtle signals, which aids models in learning from imbalanced data.

(V) Fourier transform: This is widely used in signal processing and image processing. It transforms a time-domain signal into its frequency-domain representation, allowing for analyzing and manipulating the signal's frequency components. In rare event prediction, it can be used to uncover frequency-domain characteristics that reveal rare occurrences often hidden by time-domain noise. In audio-based rare event research, [36], fourier transform is applied to convert audio waveforms into the frequency domain for further analysis.

In this section, we explored four primary techniques, data cleaning, feature selection, sampling, and feature engineering, as approaches to data processing. Data cleaning methods help to ensure data quality and remove noise, while feature selection techniques aim to identify the most relevant features for rare event prediction. Sampling approaches assist in selecting a subset of datasets and address issues with class imbalance and large dataset sizes. Feature engineering methods effectively extract meaningful information and represent the data more discriminatively. We also examined their application in various data modalities, rarity levels, and downstream tasks (See Figure 7 and discussion in Appendix B.1). It is observed that many studies utilize standard data processing techniques common to general ML research. This highlights the need to investigate standardized approaches adapted to rare events to assist the unique challenges of classification, clustering, regression, and forecasting. Notably, numerical data cleaning techniques like data sifting, filtering, imputation, and noise removal are commonly used, while feature selection methods such as correlation-based, wrapper-based, and filter-based techniques are applied across different rarity groups, with classification tasks being the primary focus across all methodologies.

4 Algorithmic Approaches

Algorithmic approaches are pivotal in any ML pipeline and contribute significantly in making informed and effective decisions. They provide mathematical models for varied use cases centered around downstream tasks like classification, clustering, forecasting, regression, and simulation. Firstly, this section analyzes a subset of algorithmic approaches utilized in the literature on rare events. Then, each approach would be examined concerning several algorithmic indicators. Finally, each approach would be analyzed with the data modality, rarity groups, data processing techniques, and downstream tasks.
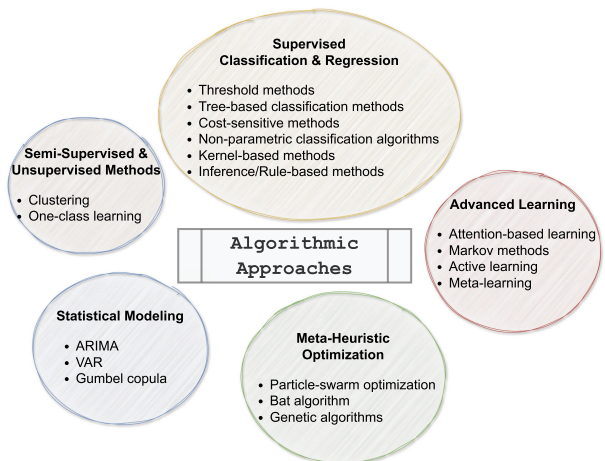

Fig. 5. Algorithmic approaches used in rare event prediction.

4.1 Significance of Algorithmic Approaches

Algorithmic approaches provide tools and modeling techniques to analyze and interpret complex datasets, enabling the identification of patterns, relationships, and key factors and predicting events of significant importance. We categorize algorithmic approaches in rare event research into five major groups: Supervised classification and regression, Clustering, Statistical modeling, Meta-heuristic optimization, and Advanced learning methods as shown in Figure 5.

4.2 Supervised Classification and Regression

Classification and regression are two fundamental supervised learning tasks in ML aimed at predicting the output of a target variable based on input features. The key difference lies in the type of target variable they handle. Classification algorithms are used to predict discrete values such as gender, binary labels (true/false), or categories like spam or not spam, while regression algorithms are employed for predicting continuous values like price, salary, or age. In rare event prediction, various techniques have been used, including threshold methods, tree-based classification, one-class learning, cost-sensitive methods, non-parametric classification algorithms, kernel-based methods, and inference/rule-based methods.

4.2.1 Threshold Methods. These approaches set a specific threshold for classifying data instances into rare or non-rare classes based on a pre-defined criterion, rendering them well-suited for modeling rare events. Instances above the threshold are considered rare, while those below it are classified as non-rare. Existing research has employed probabilistic statistical methods such as **Logistic Regression (LR)**, **Naive Bayes classifiers (NB)**, and **Neural Networks (NN)** that generate a score or probability threshold. If the categorization is binary, this probability is subsequently transferred and mapped to a binary mapping, like malignant or benign, spam or not spam, normal or abnormal. In rare event prediction, certain researchers have utilized LR as a fundamental base classifier [186, 211], while others have explored various adaptations and variations, including incorporating regularization techniques [41, 45] and utilizing weighting methods [127]. It was observed that the conventional LR encountered convergence challenges in scenarios with limited sample sizes or rare events in the data. To address this issue, an alternative approach known as Firth's logistic regression [154] has been employed, wherein a penalty is added to the log-likelihood function of the LR model. Several studies that include rare events [45, 146] have demonstrated improved performance using Firth's logistic regression. LR methods are easy to implement, interpretable, and suitable for rare event prediction with small sample sizes.

Bayes classifiers and neural networks are also threshold-based models utilized in rare event prediction research. Bayes classifiers have been used in rare event prediction by [120, 157]. Studies have explored the application of neural networks, including Deep neural networks [38, 46, 120], **Convolutional Neural Networks (CNN)** [46, 51, 119, 150, 205], **Multi-Layer Perceptron (MLP)** [36], Autoencoders [45, 133, 202], CNN-based autoencoders [51, 97, 133, 159, 202], **Long Short-Term Memory (LSTM)** autoencoders [97, 159] in various applications such as mineral prospectivity prediction [150, 205], black-swan event prediction [51], rare sound classification in audio forensics [36] and manufacturing failure prediction [97, 159]. LSTMs can preserve temporal properties of rare events by capturing sequential dependencies and managing long-term relationships through gating mechanisms. NB offers robustness, probabilistic inference, and flexibility compared to LR, while NN excels in representation learning and scalability. But, both risks being affected by the insufficient sample size in rare events [146].

4.2.2 Tree-based Classification Methods. These are supervised ML methods that partition the training data into subsets using a series of conditional statements. These splits create a tree-like

structure, each representing a logical test on a specific feature. The final model comprises multiple such trees, enabling predictions and offering insights into the relationships within the data. Tree-based classification methods are widely used in rare event prediction due to their ability to handle complex and non-linear relationships in the data. **Random Forest (RF)** [46, 63, 73, 82, 96, 132, 157] and Boosted Classification Trees, such as XGBoost [45, 73, 96, 157], are popular choices in this group. In rare event prediction, researchers have applied these models to estimate the probability of rare events like suicide attempts [63], APS failures [82, 157], and manufacturing faults [73, 96]. Data augmentation techniques have been used to enhance the performance of tree-based models further, optimizing parameters to avoid overfitting [45]. Tree-based methods also benefit from handling overfitting using random subsets during model training. However, they can be challenging to interpret due to the large number of trees in the model, making it difficult to understand the combined effect of all trees. Despite this limitation, tree-based classification methods are valuable tools for capturing complex relationships and making accurate predictions in rare event prediction tasks.

4.2.3 Cost-sensitive Learning. These methods consider the costs associated with prediction errors and other potential costs during the training of a ML model. Instead of maximizing accuracy, the focus shifts to minimizing overall misclassification costs, where each class or instance is assigned a specific misclassification cost. False negatives (misses) are assigned higher costs than false positives (false alarms). Two main approaches to cost-sensitive learning are decision trees and weighting. Decision trees employ a parameterized threshold mechanism to dynamically adjust the decision boundary of the classifier, making them suitable for modeling rare events in a manner that is nuanced and contextually adaptable [113, 120, 145, 211]. On the other hand, weighting assigns higher weights to the minority class to penalize misclassifications of the rare class [59, 95]. Researchers have explored cost-sensitive learning methods with various classifiers such as LR, RF, and **Support Vector Machines (SVM)**. Weighted Random Forest and AdaClassWeight are examples of algorithms that adaptively adjust the weights of the rare class to address the imbalanced data problem [59, 95]. Overall, it is observed that cost-sensitive learning methods offer effective solutions for handling imbalanced datasets and predicting rare events in various real-world applications.

4.2.4 Non-parametric Classification Algorithms. They are ML algorithms that do not make explicit assumptions about the data's underlying probability distribution or functional form. Unlike parametric classification algorithms, which assume a specific functional form (linear or polynomial) and estimate parameters, non-parametric algorithms learn directly from the data without assuming any specific model structure. **K-nearest neighbors (k-NN)** is an instance adopted by [157] in APS failure prediction. They are often more flexible and can capture complex relationships in the data, which becomes more advantageous in rare event prediction [119, 157].

4.2.5 Kernal-Based Methods. They are ML algorithms that transform data into a higher-dimensional space using kernel functions, enabling them to capture non-linear relationships and solve downstream tasks like classification and regression. SVMs are a well-known kernel technique that outperforms neural networks in some rare event research, particularly for small to medium datasets demanding explainable outcomes. SVM-based kernels can disambiguate hard-to-classify rare event datasets using soft-margins [40, 100, 157]. One-class SVM has been used in evaluating large sample sizes in unsupervised learning environments [41, 100]. In imbalanced healthcare data, **L.SVM (Support Vector Machine with Linear kernel)** and **R.SVM (Support Vector Machine with Radial kernel)** have been used as base classifiers for incident detection [211]. SVMs have also been applied in importance sampling for rare event detection to identify multiple

failure regions and estimate the structural failure probability of rare events [122]. **Granular SVM (GSVM)** is a variation of SVM that combines statistical learning theory and granular computing theory [180]. In [181], GSVM has been used with under-sampling techniques to improve efficacy by locally reducing redundant data through parallel processing. **Rare event weighted kernel logistic regression (RE-WKLR)** [129] is an algorithmic SVM enhancement optimized for unbalanced and rare event data. It offers the advantages of weighing, bias correction, and the strength of kernel approaches, especially when datasets are imbalanced or not linearly separable.

4.2.6   Inference/Rule-Based Methods. This class of algorithms focuses on deriving knowledge and insights from data through explicit if-then rules. In rare event literature, these methods can be Bayesian methods, Inductive algorithms, Two-phase rule induction, Knowledge-based and human interaction-based approaches, and Association rule mining. These methods are applicable to rare event prediction due to their ability to leverage expert domain knowledge, create human-readable rules, and provide interpretable outputs, which are crucial for understanding the decision-making process in scenarios where data scarcity and complexity make traditional modeling approaches challenging.

(i) Bayesian methods: These are statistical techniques that use prior information on a certain population, and they rely on Bayes' theorem to update the probability of a hypothesis based on new evidence. Bayesian methods offer multiple benefits, including handling incomplete and noisy data, understanding causal relationships between variables, and integrating domain expertise and data through Bayesian networks. Utilizing Bayesian methods, as demonstrated in [60], along with Bayesian networks, results in superior performance in forecasting daily ozone states by incorporating expert knowledge and historical data. Bayesian networks also prove useful for extreme rare event identification, as shown in [144], where the Jaynes inferential approach is used for feature engineering and optimal binning, leading to reduced features and improved identification of relevant diagnostic features.

(ii) Inductive algorithms: In rare event prediction, some researchers have proposed various inductive bias methods such as **Maximum Specificity Bias** (**MSB**) and **Instance-Based Learning** (**IBL**). MSB aims to discover specific rules for individual training examples, enhancing the performance of small disjuncts but leading to worse overall performance. IBL, also known as lazy learning, is a ML approach that focuses on the local generalization of instances based on similarity measures. **1-NN** (**1-Nearest Neighbor**) algorithm is a type of IBL used in [113] along with one-sided selection in detecting oil spills. This study shows promising results for improving accuracy in small disjuncts, but even this method has not been able to provide conclusive evidence.

(iii) Two-phase rule induction: Is a ML technique that involves a two-step process for inducing rules from data and is commonly used in data mining and knowledge discovery. The PNrule algorithm [39] is a two-phase rule induction approach that involves the discovery of positive rules (P-rules) to predict the presence of a class and negative rules (N-rules) to predict its absence. P-rules are learned in the first phase to capture the most positive cases while maintaining a respectable false positive rate. N-rules are discovered in the second phase to reduce false positives introduced by the union of P-rules while maintaining an acceptable detection rate. The study by [105] uses the PNRule method to address the challenge of maximizing recall and precision in rare event prediction. By creating rules with great accuracy, the initial phase concentrates on recall. The second phase concentrates on precision by utilizing rules that eliminate false positives from the records covered by the first phase.

(iv) Knowledge-based and human interaction-based approaches: Knowledge and human interactions have been researched and explored in predicting rare events like international conflicts, wars, coups, revolutions, massive economic depressions, and economic crises [110].

Knowledge-driven models are particularly useful in areas with limited exploration data or that have not been extensively studied. They rely on the expertise of professionals to make decisions, but they can be subjected to limitations due to their subjective nature. Some instances of using knowledge include modeling knowledge from geological experts in mineral prospectivity prediction [150, 205] and using expert knowledge like chemical equations and hypotheses in solar flare forecasting [60]. Nevertheless, while these approaches may demonstrate efficacy when applied to small and straightforward systems, their effectiveness might be low when dealing with more complex and diverse systems [67].

(v) Association rule mining: These utilize basic If/Then statements to reveal relationships between independent relational or other data repositories. [189] adopts an association rule mining approach to identify frequently occurring patterns preceding target rare events, which are subsequently integrated into a rule-based predictive model. "PREVENT" is a general purpose inter-transaction association rules mining algorithm that uses inter-transactional patterns to predict rare events in transactional databases [49]. FP-Growth, a state-of-the-art algorithm for classical association rule mining, is utilized there. The algorithm's computational cost is minimal as it involves limited scans of the database, making it well-suited for the requirements of rare event prediction.

4.3 Semi-supervised and Unsupervised Methods

Semi-supervised methods for rare event prediction leverage a combination of labeled data from the rare event class and unlabeled data from normal instances, aiming to improve predictive accuracy and generalization. In contrast, unsupervised methods rely solely on unlabeled data, focusing on identifying patterns or anomalies that deviate significantly from the norm, which can be indicative of rare events.

4.3.1 Clustering Methods. Clustering is a type of unsupervised learning where samples are categorized based on their resemblance to neighboring data points. Several clustering-based methods have been employed in the literature on rare events. Distance-based unsupervised methods such as RF clustering, **Partition Around Medoids (PAM)**, K-means, and hierarchical clustering are commonly used [118, 202]. For instance, hierarchical clustering has been used to detect abnormal production behaviors in the paper manufacturing industry [202]. K-means clustering has been utilized to identify rare change events based on the distance between common features [89]. Additionally, clustering-based undersampling techniques, like **Clustering Large Applications (CLARA)** [145] and Classification using lOcal clusterinG (COG) algorithms [198], have aimed to generate balanced sub-classes for classification. In some cases, ensembles of clustering methods, such as RF Clustering and PAM, are been combined to optimize the clustering process [147]. Moreover, the combination of nearest neighbor and **Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)** clustering with dynamic markov chains has shown promise in detecting rare events in spatiotemporal environments using sensor and traffic data [134].

4.3.2 One-class Learning. One-class learning is an unsupervised learning strategy used for extremely skewed class distributions, with the classifier being trained exclusively on data from one class. It can be considered a semi-supervised or unsupervised approach, depending on how it is implemented and the specific context of its application. Adaptation of one-class classification algorithms for imbalanced classification has been researched in early studies, and they have also been employed in rare event prediction. HIPPO (i.e., a classification method based on Hippocampus functioning) [103] is a standard method where only the rare class is learned, and **Repeated Incremental Pruning to Produce Error Reduction (RIPPER)** [62] is a standard method where the algorithm selects the majority class as its default class and learns the rules for detecting the minority class. Hamaguchi et al. [89] have proposed a variational autoencoder-based method to

learn disentangled representations on only low-cost negative samples of image data. As a result, rare events were detected as outliers. Some autoencoder models we presented under threshold methods also do intersect with this group, such as [159], which was trained only on normal samples, and [45] where samples of rare events were used for training. The advantages of one-class classifiers come at the cost of ignoring all available information about one class; consequently, this solution should be approached cautiously, as it may not be suitable for all circumstances.

4.4 Statistical Modeling

Statistical models are algorithms that apply statistics and mathematics concepts to generate a representation of data, which is then analyzed to determine any relationships or discover insights. Statistical approaches like **Autoregressive Integrated Moving Average (ARIMA)** modeling and **Vector Autoregression (VAR)** have been used in rare event prediction [41, 97]. ARIMA is an autoregressive statistical model that predicts future values based on past values. Alestra et al. have utilized ARIMA modeling separately for each aircraft to predict the degradation behavior over time [41]. Vector Autoregression follows a stochastic process model that captures linear interdependencies among multiple time series using a linear function of past values of each variable. [97] explored VAR in evaluating the LSTM-Autoencoder model developed. ARIMA and VAR models effectively capture time-based dependencies and sequential (temporal) data patterns, which enhances prediction accuracy and addresses challenges in predicting rare events based on past observations. The Gumbel copula function is a statistical modeling technique in copula-based algorithms utilized in rare event research. Copulas are mathematical functions used to model the dependence structure between random variables. They describe the joint distribution of variables by connecting their marginal distributions. The Gumbel copula function, specifically, is a type of copula that is based on the Gumbel distribution. It is commonly used to model extreme-value dependence, making it suitable for applications involving rare events or extremes. It captures the tail dependence between variables, which is vital in scenarios where the behavior of variables in the extreme tails is of interest [206].

4.5 Meta-Heuristic Optimization

Meta-heuristic algorithms are optimization techniques for solving complex problems by iteratively exploring and exploiting the search space to find near-optimal solutions [37]. These algorithms, such as particle swarm optimization and the bat algorithm, have been applied to improve efficiency and scalability on large imbalanced healthcare data, [121], and genetic algorithms have been utilized for learning to predict rare events in categorical time series data [195]. Additionally, evolutionary ensemble algorithms have shown promise in enhancing the identification rate of minority classes in rare event-based imbalanced datasets [112]. These algorithms would excel in rare event prediction due to their ability to efficiently explore complex search spaces and adaptively exploit the underlying patterns and structures in imbalanced datasets.

4.6 Advanced Learning Methods

Recently, advanced learning methods like Attention-based mechanisms, Markov methods, Active learning, and Meta-learning have emerged to tackle the challenges posed by rare event datasets. These approaches strive to enhance rare event predictive capability by leveraging temporal dependencies, probabilistic modeling, and selective data labeling.

4.6.1 Attention-based Mechanisms. Attention mechanisms enable the model to selectively concentrate on important input elements for accurate predictions while disregarding less relevant parts. This has gained attention in rare event research as well. For instance, Liu et al. proposed a

fault diagnosis approach using one-dimensional CNN, **Gated Recurrent Unit (GRU)**, and attention mechanisms, combined with knowledge graphs, to achieve more precise predictions in bearing fault detection [123]. Xu et al. have utilized attention-based-LSTM and Extra-Tree models for fault mode and severity prediction based on bearing datasets, while Ravindranath et al. introduced the M2NN architecture with an attention mechanism for post-traumatic seizure detection, demonstrating its effectiveness in finding unusual seizures in EEG data [162, 203]. Kulkarni and team utilized soft-attention CNN for rare marker event detection and localization in well-logs [115]. Attention-based mechanisms preserve temporal properties by selectively focusing on important time steps, capturing both short- and long-term dependencies in sequential data. These methods have proven useful in handling rare event detection tasks, such as anomaly detection in manufacturing, rare disease diagnosis, and identifying specific rare occurrences in multiple domains.

4.6.2   Markov Methods. Markov models represent the Markov property, which asserts that the prediction of an outcome depends only on information about the current state and is independent of the previous sequence of events. **Extensible Markov Models (EMM)** and Monte Carlo methods are among the markov methods adopted in rare event literature. These models capture the time-dependent behavior by modeling transitions between states over time and, hence, can address the temporal properties inherent in rare events.

(I) Extensible markov models (EMM): EMM are a dynamic variation of traditional static Markov models. They excel in modeling spatiotemporal data and have proven to help predict spatiotemporal events, including rare occurrences, by capturing spatial, temporal, and unusual event transitions [72]. Meng et al. [134] introduced an approach that combines clustering and EMM for rare event detection in spatiotemporal data, particularly sensor and traffic data. Their contribution lies in modeling the transitions between rare and typical events using EMM, effectively detecting rare occurrences.

(II) Monte carlo methods: Monte carlo methods, also called Monte carlo simulations, are probabilistic mathematical techniques that estimate the possible outcomes of uncertain events. In rare event estimation, Monte Carlo methods have been applied to estimate model uncertainty, such as using Monte Carlo Dropout to create multiple predictions and measuring the standard deviation of detection depths as a proxy for uncertainty [115]. Additionally, they have been used to predict rare events using trajectory data and unscheduled aircraft maintenance actions [28, 46, 66, 67, 122, 146, 179, 196].

4.6.3   Active Learning. Active learning is a learning strategy that iteratively chooses the most instructive or uncertain instances from a pool of unlabeled data and asks an oracle (such as a human expert or a pre-existing labeled dataset) to annotate their labels [171]. Recent research efforts have been to predict rare events using active learning, exemplified by studies like [70, 152]. Dhulipala et al. [70] proposed a methodology for rare event simulation by combining active learning and multi-fidelity modeling, which leverages multiple levels of models' fidelity (accuracy or complexity) to predict outcomes efficiently. **Deep Neural Operators (DNOs)**, like DeepONet [125], are nonlinear operators created to handle systems with infinite dimensions, making them highly effective surrogate models for precisely representing extreme events. Pickering et al. [152] introduced a Bayesian-inspired framework based on active learning for DNOs, specifically designed for discovering and quantifying extreme events, focusing on uncertainty quantification.

4.6.4   Meta Learning. Meta-learning, or "learning to learn", is a ML approach focusing on training models to effectively adapt and generalize to new tasks or domains with minimal data. It can be applied to rare event prediction by training a model on various rare event scenarios. With limited data, it can rapidly adapt and accurately predict new and infrequent events. The advantage

Table 2. Comparison of Algorithmic Approaches

Major Factor	Algorithmic Indicator	Supervised Classif. & Regress.	Semi-Sup. & Un-Sup.	Statistical Modeling	Meta-Heuristic Optimization	Advanced Learning Methods
Computational Efficiency	Training Time	✓*	✓	✓	✓	✓*
	Memory Usage	✓*	✓	✓	✓	✓*
	Model Size	✓*	✓	✓	✓	✓*
	Model Complexity	✓*	✓*	✓*	✓*	✓*
Model Analysis & Understanding	Feature Importance	✓	-	✓	-	✓
	Model Explainability	✓	-	✓	-	✓
	Interpretability	✓	-	-	-	✓
Data Availability	Labeled Data	✓	✓	✓	✓	✓
	Unlabeled Data	-	✓	-	-	✓
Model Performance	Performance on Large Data	✓	-	✓	✓*	✓
	Ability to Handle Noise	✓	-	✓	✓*	✓
	Generalization	✓	✓	✓	✓	✓

∗ includes variability- depends on specific factors such as the problem domain, dataset, or implementation.

of meta-learning over other advanced learning approaches lies in its ability to dynamically adapt and integrate different predictive techniques, a faster and cheaper training process, thus enhancing the model's generalization capability and performance across varied scenarios. Several studies [88, 117, 173] have looked at how meta-learning techniques can be used to improve failure prediction in different areas, such as large-scale computing systems and acoustic event detection, by dynamically combining base methods and improving accuracy in situations with few labeled data. While this existing research primarily focuses on failure prediction, it is worth noting that no prior studies have specifically addressed the prediction of rare events using meta-learning techniques.

4.7 Comparison of Algorithmic Approaches

Table 2 offers a comparative analysis of the approaches discussed in the preceding section. We use 12 algorithmic indicators based on four major factors: computational efficiency, model analysis and understanding, data availability, and model performance, as defined in Appendix C.1. Table 7 (Appendix C.2) analyses algorithmic approaches with rarity groups, downstream tasks, modalities, dataset types, and data processing tasks based on our reviewed papers. Predominantly, research has centered on classifying numeric rarity within real datasets, showcasing a blend of various data processing methodologies. Limited research is dedicated to alternative downstream tasks such as clustering, forecasting, simulation, and diverse modalities encompassing text, images, and audio.

This section extensively examined different algorithmic approaches applied in rare event prediction. The findings contribute to a better understanding of the diverse strategies employed in rare event prediction research and highlight the importance of selecting appropriate algorithms based on specific downstream tasks and modality requirements.

5 Evaluation Approaches

As ML models are developed, measuring the performance of every model is critical. Multiple evaluation metrics are employed in ML depending on the model and the results produced. This section focuses on the varied evaluation aspects used in rare event prediction. Firstly, we will highlight the significance of evaluation concerning rare event studies, followed by the evaluation methodologies and performance metrics used in related studies.

5.1 Significance of Evaluation

Evaluation in rare event prediction research plays an integral role as it allows the assessment of the model's performance in accurately identifying and predicting rare events. Due to the imbalanced nature of rare events, traditional evaluation metrics might not provide a comprehensive assessment, making it vital to employ specialized evaluation techniques that focus on both overall performance and the ability to detect rare occurrences. Figure 6 depicts the evaluation approaches we identified in rare event literature grouped into their evaluation categories.
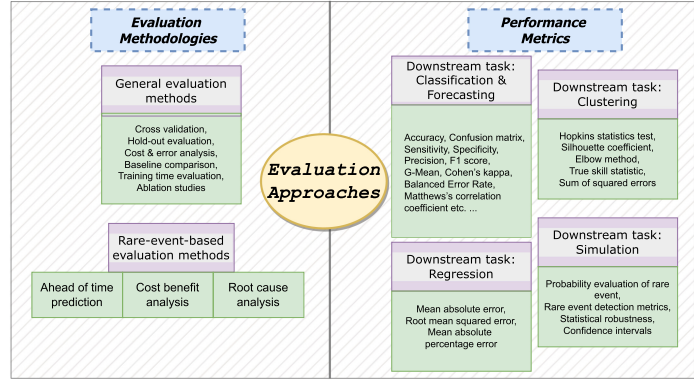

Fig. 6. Evaluation approaches in rare event prediction.

5.2 Evaluation Methodologies

Evaluation methodologies in rare event prediction research involve general and rare event-specific approaches. General methods find widespread application in diverse ML tasks, while rare event-specific metrics are tailored to address the unique challenges and requirements of predicting rare events.

5.2.1 General Evaluation Methods. In rare event prediction research, various general evaluation methods are employed to assess the performance of algorithmic approaches (Table 8 in Appendix D). Cross-validation techniques, such as K-Fold, Stratified K-Fold, and Leave-One-Out are commonly used for performance estimation [40, 42, 59, 63, 73, 82, 96, 113, 119, 161, 211]. Holdout evaluation through random splitting [63, 96, 123, 160] and time-based splitting [60, 63, 67] is utilized to create distinct train-test datasets, facilitating the process of model training and evaluation. Analysis with standard baselines, cost and error analysis, training time evaluation [51], and ablation studies [89] for understanding individual components' contributions are among the general evaluation approaches employed in this field.

5.2.2 Rare Event-specific Evaluation Methods. Evaluation methods that are specific to rare events emphasize early prediction, proactive decision-making, and economic feasibility in the prediction task. We identified several methods below as specific to rare events as presented in Table 9 (Appendix D).

(I) Ahead-of-time prediction evaluation: Ahead-of-time prediction evaluation assesses a model's predictive capability for events or outcomes occurring multiple time steps into the future. This approach contrasts with the usual prediction evaluation, which focuses on predicting the immediate next event or outcome. Evaluating models "ahead of time" is particularly important in scenarios where early detection and proactive decision-making are critical. In the studies by [51, 67, 159], the performance of the proposed models has been evaluated over various lead times, ranging from one to many units of time. It was observed that while the accuracy of the models

might decrease as the lead time increases, their performance has remained stable, demonstrating their potential utility in providing valuable insights well before the occurrence of rare events, such as black-swan or equipment failures.

(II) Cost-benefit analysis: Cost-benefit analysis has been used to assess rare event prediction's economic feasibility and potential advantages. Methods like assessing potential financial loss reduction, optimizing resource allocation, and evaluating the potential mitigation of losses or damages are utilized to quantify and compare the costs and benefits linked to early prediction of rare events, as explored in [119] and [144].

(III) Root cause analysis: This could be viewed as a problem-solving approach focused on discerning the fundamental elements contributing to rare occurrences. It involves investigating and tracing the chain of events to determine the fundamental causes [119].

5.3 Performance Metrics

Performance metrics aid in monitoring and optimizing algorithmic approaches, facilitating comparison of algorithms, determining optimal thresholds, supporting data-driven decision-making and conducting risk assessments. We categorize these performance metrics used in rare event research, which differ based on the downstream tasks involved. The detailed analysis of performance metrics and rarity groups, algorithmic approaches, modality, and dataset types is included in Table 10 (Appendix D).

(I) Downstream tasks: Classification and Forecasting

(i) Accuracy: Evaluating model accuracy is essential, particularly in imbalanced scenarios, where relying solely on accuracy can be deceptive. Since accuracy only assigns overall class weights rather than weights for unusual classes or minorities. Zhao et al. [211] highlight that standard classifiers like LR and SVM assume equal class distribution, which leads to poor sensitivity for rare events. To address this, variants of accuracy such as **Faulty-Normal Accuracy** (**FNACC**) and **Real Faulty-Normal Accuracy** (**RFNACC**) are introduced [41, 132]. FNACC assesses how well the model identifies normal samples preceding faults, while RFNACC evaluates the accuracy in predicting actual faults, offering a more nuanced view of model performance on rare events.

(ii) Geometric mean (G-Mean): G-Mean combines sensitivity and specificity to measure the balance between classification performance for the majority and minority classes. It is particularly useful in rare event prediction as a low G-Mean indicates poor prediction of rare events, despite good performance on the majority class [59, 181, 191].

(iii) Cohen's kappa index: It is a metric that has been used in earth science and healthcare-based rare event prediction evaluation for use cases like landslide prediction [186], hazardous seismic bumps in coal mines, detecting changes in geospatial trajectories [100, 120], mineral prospectivity prediction [205] and thoracic events prediction [121]. [120] has selected Kappa as the primary evaluation metric since it indicates the generalizability of the classifier's predictive ability on the supplementary datasets. However, some research attempts have proven that the Kappa metric produces unreliable results due to its high sensitivity to the distribution of the marginal totals [61].

(iv) Balanced error rate (BER): BER measures the average of the errors for each class in a classification problem, which has been used by [100, 120] in rare event model evaluation.

(v) Matthews's correlation coefficient (MCC): MCC is a more accurate statistical metric that provides a reliable measure of a model's performance by considering all four confusion matrix categories (TP, FN, TN, and FP) and balancing the impact of both positive and negative class sizes [61, 120, 204]. Hence, this metric is beneficial when dealing with highly imbalanced rare event datasets.

(vi) Cost matrix: A cost matrix specifies the relative importance of accuracy for different classification predictions by assigning costs or weights to various classification predictions. In a cost

matrix, positive numbers (costs) influence negative outcomes, whereas negative numbers (benefits) influence positive outcomes. An instance is [73], which tries to minimize the cost matrix for binary misclassification using threshold and tree-based classification models. This approach helps prioritize detecting rare events by adjusting the cost of false positives and false negatives.

(vii) Top-decile lift (TDL): TDL is a metric used in the economic domain that compares the incidence in the top 10% samples with the highest model predictions to the incidence of the entire sample [50]. [42] used TDL and assumed it returned the customers predicted to be most likely to churn rather than a random selection. In rare event scenarios, TDL helps the model identify the most promising cases by highlighting its ability to distinguish rare events from non-rare ones, thus providing insights beyond overall accuracy.

(viii) Reconstruction error: This measures the discrepancy between the original input data and its reconstructed version generated by a model by quantifying how well the model can reproduce the input data. In models like autoencoders, the model is trained to reconstruct its input data from a compressed latent representation [97]. Reconstruction error identifies significant deviations from normal patterns, aiding in effective rare event prediction and enhancing the model's ability to prioritize and identify rare events.

(ix) Moran index: Used in spatial analysis and statistics and measures a variable's spatial autocorrelation or clustering within a geographic space. It calculates the degree of similarity or dissimilarity between neighboring locations based on the values of the analyzed variable. It provides a measure of spatial dependence, indicating whether similar values tend to cluster together (positive spatial autocorrelation) or are dispersed (negative spatial autocorrelation) [206]. Moran index has been used to detect spatial clustering of rare events, and it gives insights into spatial locations where these events are more likely to occur [168, 206].

(II) Downstream task: Clustering

(i) Elbow method: The Elbow method, typically used to identify the optimal number of clusters, is beneficial in rare event prediction tasks where the majority class can dominate the clustering process. This also ensures that minority class clusters are not merged into larger majority class clusters, thereby preserving the distinct characteristics of rare events. This approach is helpful in studies focusing on class imbalance, where detecting minority classes with high specificity and sensitivity is vital for model performance [145].

(ii) Silhouette coefficient: It is another metric used for internal cluster validation. This metric is particularly important in rare event prediction as it helps to determine whether rare event clusters are well-defined and distinct from non-rare event clusters. A small silhouette width may indicate that rare events are being inaccurately grouped with non-rare events, which can significantly impact the model's ability to identify and predict rare occurrences [147].

(iii) Hopkins statistics test The Hopkins statistics test is used to assess the clustering tendency of a dataset. Omar et al. [147] used it to evaluate the clustering tendency in a rare event dataset. This revealed well-defined and meaningful clusters that standard clustering evaluation methods might miss due to the high data imbalance and the distinct patterns often exhibited by the minority class.

(iv) True skill statistic (TSS): TSS combines sensitivity and specificity into a single measure, offering a more balanced model performance evaluation in rare event prediction. Unlike accuracy, which can be misleading in imbalanced datasets, TSS provides a robust assessment of a model's positive and negative predictive abilities [43, 208]. This makes it particularly useful in scenarios like species distribution modeling or solar flare forecasting, where both false positives and false negatives carry significant consequences [40].

(v) Sum of squared errors (SSE): SSE measures the variation within clusters by calculating the sum of the squared differences between each observation and the group mean and is often used to evaluate cluster compactness [182]. In rare event prediction, minimizing SSE ensures that rare

event clusters are as distinct and well-defined as possible [165]. High SSE values may indicate that the model fails to accurately capture the unique characteristics of rare events, leading to poor prediction performance [165].

(III) Downstream task: Regression

(i) Mean absolute error (MAE) and root mean squared error (RMSE), mean absolute percentage error (MAPE):

MAE or **Mean Absolute Deviation** (**MAD**), RMSE, and MAPE or **Mean Absolute Percentage Deviation** (**MAPD**) are popular scale-dependent metrics used in evaluation. MAE provides a straightforward measure of prediction error, highlighting performance on underrepresented rare events. RMSE penalizes more significant errors, making it ideal for assessing precision in datasets where rare occurrences have high variance. Though sensitive to small actual values, MAPE offers insights into percentage accuracy, which is essential for comparing model performance across different events. These metrics have been employed in rare event studies, with Xu et al. [203] selecting MAE and RMSE as regression loss functions and Ravindranath et al. [162] using these metrics in multivariate multiscale attention research.

(IV) Downstream task: Simulation

(i) Probability evaluation of rare event: This metric directly measures the likelihood of the rare event occurring. It provides a quantitative assessment in a simulation of the occurrence of the event of interest [46, 70, 146].

(ii) Rare event detection metrics: The standard metrics that evaluate the performance of algorithms designed to predict rare events like precision, recall, F1 score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC), RMSE, have also been utilized in simulation evaluations [46, 146].

(iii) Statistical robustness: Statistical robustness refers to the ability of statistical models to produce reliable and consistent results under varying sample sizes, distributional assumptions, or data perturbations [46].

(iv) Confidence intervals: This metric provides a range of values within which the actual value of a performance metric is expected to fall and quantifies the uncertainty associated with the estimated performance metric [146].

In conclusion, we presented a comprehensive analysis of various evaluation approaches used in rare event prediction. From our analysis, we identified most studies rely on standard evaluation metrics, highlighting the need for tailored metrics explicitly designed for rare events. This is particularly important given the limited availability of rare event-specific evaluation techniques, especially in datasets comprising numerical, textual, and image data across diverse tasks such as classification, forecasting, clustering, regression, and simulation. Through this examination, we also gained valuable insights into the suitability and limitations of different evaluation methodologies and performance metrics for handling rare events. The findings underscore the significance of choosing appropriate metrics tailored to specific downstream tasks and modalities.

6 Research Findings and Discussion

This section provides an overview of the research findings, emphasizing identified gaps, and open challenges in the field. It also explores emerging research trends and novel approaches, aiming to address the complexities of rare event prediction and drive advancements in the field.

6.1 The Gaps in Current Literature

(1) Lack of standardized benchmark datasets: "Standardized dataset" represents real-world scenarios that include an ideal rarity percentage, annotated ground truth, scalability,

temporal and spatial association, feature diversity, noise, and outliers. They should also adhere to privacy and ethical standards to enable fair and comprehensive comparisons of algorithms and advance the state-of-the-art in the field. The absence of such widely accepted benchmark datasets for rare event prediction makes it challenging to compare the performance of different algorithms and approaches consistently.

(2) Imbalanced data processing techniques: While many studies explore standard methods for addressing imbalanced data, there is a necessity for in-depth investigation of standardized techniques for data processing, especially for handling extreme rarity, since general data imbalance techniques may inadequate and underperform.

(3) Scalability issues: Current research has not addressed scalability challenges in handling large-scale rare event datasets, such as computational complexity, memory constraints, and algorithmic efficiency.

(4) Limited focus on **Uncertainty Quantification** (**UQ**): UQ involves quantifying and characterizing uncertainties associated with predictions, offering insights into the reliability and confidence of estimates. The lack of comprehensive studies on UQ for rare event prediction undermines confidence in model predictions, which is essential for making informed decisions when dealing with infrequent and high-impact events.

(5) Real-world applicability: More research is needed to assess the practical usability and robustness of prediction models in critical real-world applications, particularly in domains such as healthcare, finance, and earth sciences.

(6) Limited focus on rare event-specific evaluation techniques: Many studies still rely on standard and general evaluation metrics designed for balanced datasets, which is challenging because these metrics fail to accurately reflect model performance in rare events, leading to misleading assessments and suboptimal model improvements.

(7) Measuring data quality: The absence of standardized methods for measuring data quality in data sampling and augmentation for rare event data is a research gap.

6.2 Open Challenges of Rare Event Prediction

Based on our study, below are some open issues and challenges in the field of rare event prediction that we identified.

(1) Lack of annotated data: Collecting and annotating data about rare events proves complex and time-intensive, posing an inherent obstacle in developing proficient rare event prediction models.

(2) Interdependence between rare events: Identifying complex relationships and dependencies among rare events (when more than one rare event occurs) remains elusive.

(3) Bias in data: Bias refers to systematic errors or prejudices in data collection, sampling, or labeling that can skew the results of prediction models. This bias is especially challenging in rare events since it can lead to significant misrepresentations of the events, resulting in highly inaccurate and unreliable predictions.

(4) Generalization and real-world applicability: Achieving generalizability in rare event prediction models presents challenges due to the potentially significant differences in variability and distribution of new, unseen, or real-world data compared with the training data.

(5) Interpretability: The infrequency and high stakes of rare events necessitate clear and understandable explanations to ensure trust and accurate decision-making in critical domains like healthcare or finance.

6.3 Research Trends in Rare Event Prediction

Research in rare event prediction has surged in recent years due to the increasing importance of rare events in many fields. Some of the research trends in this area include:

(1) Incorporation of domain knowledge: Domain knowledge, expert opinions, and human insights Human-in-the-Loop Systems) can aid in rare event prediction. They have led in emerging applications that combine data-driven and algorithmic approaches with qualitative knowledge, knowledge graphs, expert systems, and rule-based models to improve the robustness and reliability of predictions [79, 135, 169, 188, 199].

(2) Explainability: The growing emphasis on transparency and comprehension in model predictions has led to the rise of explainability as a prospective future trend in rare event prediction research [74, 166, 166]. This development aims to enhance decision-making processes and foster trust in the accuracy of models.

(3) Ensemble learning: This involves combining multiple algorithms to improve prediction performance, which is particularly effective in handling imbalanced datasets and has recently gained attention [69, 76, 131, 138].

(4) Meta-learning, few-shot learning, and transfer learning: Meta-learning [151] and few-shot learning [209] have proven to enhance rare event prediction recently by enabling model learning from related events and generalization to new and infrequent events with limited labeled data, thus improving adaptability and accuracy. Transfer learning is a meta-learning technique that involves using knowledge from one domain to improve learning in another. This approach is promising in rare event prediction, where data from different domains can be leveraged to improve predictive performance [69, 101].

(5) Using multi-modal data: Integrating information from multiple data sources and modalities has improved rare event predictive power, generalization, and robustness while uncovering hidden patterns [44, 163]. However, it also presents challenges related to data integration, feature engineering, and model complexity.

(6) Uncertainty quantification with rarity: Recently, UQ has gained prominence within rare event prediction. Using UQ, decision-makers can make more informed choices, assess risk, optimize resource allocation, and enhance the robustness to mitigate the impacts of rare events. These can incorporate advancements using probabilistic modeling [199], Bayesian inference [83], Monte Carlo methods [86], and Deep Ensembles [86, 108].

(7) Privacy-preserving techniques: Rare event research on techniques like federated learning, differential privacy, encrypted data processing [55, 102], and privacy-preserving data mining [87] can be prioritized for sensitive rare event data to ensure confidentiality .

(8) Leveraging edge devices and neuromorphic computing techniques: Incorporating edge devices and advanced neuromorphic computing techniques has emerged as a trend in rare event prediction [158, 172], enabling efficient real-time processing, low-latency analytics, and enhanced model adaptability at the edge while leveraging efficient computing paradigms for improved pattern recognition and prediction accuracy.

We have also formulated a forward-looking vision to assist researchers in navigating future initiatives within this field, emphasizing three pivotal areas of interest: causality, process knowledge workflows and automated planning (See description and Figure 11 in Appendix E.

7 Conclusion

This article presents a detailed analysis of rare event prediction, encompassing rare event data, data processing, algorithmic and evaluation approaches. Through our examination of the current

literature, we identified several gaps and challenges in the field, highlighting the need for specialized evaluation techniques and their integration. As rare events continue to play a vital role in diverse domains, like manufacturing, healthcare, finance, and earth sciences, addressing these challenges will foster the development of more robust and accurate prediction models. By embracing emerging research trends and leveraging advanced learning methods, we can unlock new opportunities to enhance the prediction and management of rare events, ultimately contributing to safer and more efficient decision-making processes. As the field evolves, interdisciplinary collaboration and innovative solutions will pave the way for transformative advancements in rare event prediction.

8 Acronyms

Table 11 under the Appendix F includes the acronyms used in the article.

REFERENCES

[1] 1975. Access to Forecasts - ecmwf.int. Retrieved May 14, 2023 from https://www.ecmwf.int/en/forecasts/accessing-forecasts

[2] 1987. UCI Machine Learning Repository: Data Sets — archive.ics.uci.edu. Retrieved May 13, 2023 from https://archive.ics.uci.edu

[3] 1996. Adult Dataset. Retrieved May 09, 2023 from https://archive.ics.uci.edu/ml/datasets/adult

[4] 1998. UCI Machine Learning Repository: KDD Cup 1999 Data Data Set — archive.ics.uci.edu. Retrieved May 14, 2023 from https://archive.ics.uci.edu/ml/datasets/kdd+cup+1999+data

[5] 1999. KDD Cup 1999 Data — kdd.ics.uci.edu. Retrieved May 14, 2023 from http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[6] 1999. UCI Machine Learning Repository: Spambase Data Set — archive.ics.uci.edu. Retrieved May 09, 2023 from https://archive.ics.uci.edu/ml/datasets/spambase

[7] 2004. KEEL: A Software Tool to Assess Evolutionary Algorithms for Data Mining Problems (Regression, Classification, Clustering, Pattern Mining and so on) — sci2s.ugr.es. Retrieved May 13, 2023 from https://sci2s.ugr.es/keel/datasets.php

[8] 2005. Google Maps Platform | Google Developers — developers.google.com. Retrieved May 14, 2023 from https://developers.google.com/maps

[9] 2007. UCI Machine Learning Repository — archive.ics.uci.edu. Retrieved Jun 13, 2023 from https://archive.ics.uci.edu/dataset/159/magic+gamma+telescope

[10] 2008. Aircraft Condition Monitoring System (ACMS) | SKYbrary Aviation Safety - skybrary.aero. Retrieved May 07, 2023 from https://www.skybrary.aero/articles/aircraft-condition-monitoring-system-acms

[11] 2009. PHM Society — phmsociety.org. Retrieved August 05, 2023 from https://phmsociety.org/

[12] 2010. Find Open Datasets and Machine Learning Projects | Kaggle — kaggle.com. Retrieved May 13, 2023 from https://www.kaggle.com/datasets

[13] 2013. UCI Machine Learning Repository: Data Set — archive.ics.uci.edu. Retrieved May 07, 2023 from http://archive.ics.uci.edu/ml/datasets/seismic-bumps

[14] 2016. IEEE 39-Bus System — electricgrids.engr.tamu.edu. Retrieved from https://electricgrids.engr.tamu.edu/electric-grid-test-cases/ieee-39-bus-system/

[15] 2017. ABCD Dataset - datasets.abci.ai. Retrieved May 14, 2023 from https://datasets.abci.ai/dataset/abcd/

[16] 2017. GitHub - gistairc/ABCDdataset — github.com. Retrieved May 14, 2023 from https://github.com/gistairc/ABCDdataset

[17] 2017. UCI Machine Learning Repository: APS Failure at Scania Trucks Data Set — archive.ics.uci.edu. Retrieved May 14, 2023 from https://archive.ics.uci.edu/ml/datasets/APS+Failure+at+Scania+Trucks

[18] 2018. GitHub - MHResearchNetwork/Diagnosis-Codes: Information on Diagnosis Code Lists used in MHRN, Including Formats and Datasets — github.com. Retrieved May 07, 2023 from https://github.com/MHResearchNetwork/Diagnosis-Codes

[19] 2018. KDD Cup 1999 Data — kaggle.com. Retrieved April 19, 2023 from https://www.kaggle.com/datasets/galaxyh/kdd-cup-1999-data

[20] 2019. GitHub - subhande/APS-Failure-at-Scania-Trucks-Data-Set — github.com. Retrieved August 05, 2023 from https://github.com/subhande/APS-Failure-at-Scania-Trucks-Data-Set

[21] 2020. 5 year BSE Sensex Dataset-kaggle.com. Retrieved May 07, 2023 from https://www.kaggle.com/datasets/ravisane1/5-year-bse-sensex-dataset

[22] 2020. Stroke Prediction Dataset — kaggle.com. Retrieved May 09, 2023 from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[23] 2021. Bearing Data Center: Case School of Engineering: Case Western Reserve University. Retrieved May 09, 2023 from https://engineering.case.edu/bearingdatacenter

[24] 2021. NIFTY-50 Stock Market Data (2000-2021) — kaggle.com. Retrieved May 07, 2023 from https://www.kaggle.com/datasets/rohanrao/nifty50-stock-market-data

[25] 2022. Audio-Anomaly-Dataset - kaggle.com. Retrieved May 09, 2023 from https://www.kaggle.com/datasets/ahmedabbasi/audioanomalydataset

[26] 2022. Cerberus Web Client — ftp.fdot.gov. Retrieved May 17, 2023 from https://ftp.fdot.gov/file/d/FTP/FDOT/co/planning/transtat/gis/s hs_maps/shsmap.pdf

[27] 2022. Cerberus Web Client — ftp.fdot.gov. Retrieved May 17, 2023 from https://ftp.fdot.gov/file/d/ FTP/FDOT/co/planning/transtat/gis/TRANSTAT_metadata/rdaccess.shp.xml

[28] 2022. GitHub - Chao-Dang/Rare-Event-Estimation-by-Parallel-Adaptive-Bayesian-Quadrature — github.com. Retrieved May 15, 2023 from https://github.com/Chao-Dang/Rare-Event-Estimation-by-Parallel-Adaptive-Bayesian-Quadrature

[29] 2022. GitHub - petrobras/3W: This is the First Repository Published by Petrobras on GitHub. It Supports the 3W Project and Promotes Experimentation of Machine Learning-based Approaches and Algorithms for Specific Problems Related to Undesirable Events that Occur in Offshore Oil Wells. — github.com. Retrieved May 09, 2023 from https://github.com/petrobras/3W

[30] 2022. Intelligent Systems Division — nasa.gov. Retrieved May 15, 2023 from https://www.nasa.gov/intelligent-systems-division

[31] 2022. NASA Open APIs — api.nasa.gov. Retrieved May 13, 2023 from https://api.nasa.gov/

[32] 2023. OLGA Dynamic Multiphase Flow Simulator — software.slb.com. Retrieved May 15, 2023 from https://www.software.slb.com/products/olga

[33] 2023. Roadway Data Home - TDA, MnDOT — dot.state.mn.us. Retrieved May 07, 2023 from https://www.dot.state.mn.us/roadway/data/[Accessed 07-May-2023].

[34] 2023. S and P BSE SENSEX Dataset-bseindia.com. https://www.bseindia.com/

[35] 2023. Storm Prediction Center Storm Reports — spc.noaa.gov. Retrieved May 14, 2023 from https://www.spc.noaa.gov/climo/

[36] Ahmed Abbasi, Abdul Rehman Rehman Javed, Amanullah Yasin, Zunera Jalil, Natalia Kryvinska, and Usman Tariq. 2022. A large-scale benchmark dataset for anomaly detection and rare event classification for audio forensics. *IEEE Access* 10 (2022), 38885–38894.

[37] Mohamed Abdel-Basset, Laila Abdel-Fatah, and Arun Kumar Sangaiah. 2018. Metaheuristic algorithms: A comprehensive review. *Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications* (2018), 185–231.

[38] Syed M. Adil, Cyrus Elahi, Dev N. Patel, Andreas Seas, Pranav I. Warman, Anthony T. Fuller, Michael M. Haglund, and Timothy W. Dunn. 2022. Deep learning to predict traumatic brain injury outcomes in the low-resource setting. *World Neurosurgery* 164 (2022), e8–e16.

[39] Ramesh Agarwal and Mahesh V. Joshi. 2001. PNrule: A new framework for learning classifier models in data mining (a case-study in network intrusion detection). In *Proceedings of the 2001 SIAM International Conference on Data Mining*. SIAM, 1–17.

[40] Azim Ahmadzadeh, Berkay Aydin, Dustin J. Kempton, Maxwell Hostetter, Rafal A. Angryk, Manolis K. Georgoulis, and Sushant S. Mahajan. 2019. Rare-event time series prediction: A case study of solar flare forecasting. In *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA'19)*. IEEE, 1814–1820.

[41] S. Alestra, C. Bordry, C. Brand, E. Burnaev, P. Erofeev, A. Papanov, and C. Silveira-Freixo. 2014. Rare event anticipation and degradation trending for aircraft predictive maintenance. In *Proceedings of the 11th World Congress on Computational Mechanics, WCCM*, Vol. 5. 6571.

[42] Özden Gür Ali and Umut Arıtürk. 2014. Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. *Expert Systems with Applications* 41, 17 (2014), 7889–7903.

[43] Omri Allouche, Asaf Tsoar, and Ronen Kadmon. 2006. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* 43, 6 (2006), 1223–1232.

[44] Henri Arno, Klaas Mulier, Joke Baeck, and Thomas Demeester. 2023. From numbers to words: Multi-modal bankruptcy prediction using the ECL dataset. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, Hsin-Hsi Chen, Hiroki Sakaji, and Kiyoshi Izumi (Eds.). Association for Computational Linguistics, Bali, Indonesia, 11–21. https://doi.org/10.18653/v1/2023.finnlp-2.2

[45] Md Tanvir Ashraf, Kakan Dey, and Sabyasachee Mishra. 2023. Identification of high-risk roadway segments for wrong-way driving crash using rare event modeling and data augmentation techniques. *Accident Analysis & Prevention* 181 (2023), 106933.

[46] Yuanlu Bai, Zhiyuan Huang, Henry Lam, and Ding Zhao. 2022. Rare-event simulation for neural network and random forest predictors. *ACM Transactions on Modeling and Computer Simulation* 32, 3 (2022), 1–33.

[47] Mohamed Bekkar and Taklit Akrouf Alitouche. 2013. Imbalanced data learning approaches review. *International Journal of Data Mining & Knowledge Management Process* 3, 4 (2013), 15.

[48] Zied Ben Bouallègue and David S. Richardson. 2022. On the ROC area of ensemble forecasts for rare events. *Weather and Forecasting* 37, 5 (2022), 787–796.

[49] Christos Berberidis and Ioannis Vlahavas. 2007. Detection and prediction of rare events in transaction databases. *International Journal on Artificial Intelligence Tools* 16, 05 (2007), 829–848.

[50] Michael J. A. Berry and Gordon S. Linoff. 2004. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* John Wiley & Sons.

[51] Samit Bhanja and Abhishek Das. 2022. A Black Swan event-based hybrid model for Indian stock markets' trends prediction. *Innovations in Systems and Software Engineering* 20 (2022), 1–15.

[52] Albert Bifet, Geoff Holmes, Bernhard Pfahringer, Philipp Kranen, Hardy Kremer, Timm Jansen, and Thomas Seidl. 2010. Moa: Massive online analysis, a framework for stream classification and clustering. In *Proceedings of the 1st Workshop on Applications of Pattern Analysis.* PMLR, 44–50.

[53] Bosch. 2016. Bosch Production Line Performance - kaggle.com. Retrieved April 19, 2023 from https://www.kaggle.com/c/bosch-production-line-performance

[54] Richard R. Brooks and S. Sitharama Iyengar. 1996. Robust distributed computing and sensing algorithm. *Computer* 29, 6 (1996), 53–60.

[55] Syed Muhammad Salman Bukhari, Muhammad Hamza Zafar, Mohamad Abou Houran, Syed Kumayl Raza Moosavi, Majad Mansoor, Muhammad Muaaz, and Filippo Sanfilippo. 2024. Secure and privacy-preserving intrusion detection in wireless sensor networks: Federated learning with SCNN-Bi-LSTM for enhanced reliability. *Ad Hoc Networks* 155 (2024), 103407.

[56] Ander Carreño, Iñaki Inza, and Jose A. Lozano. 2020. Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework. *Artificial Intelligence Review* 53, 5 (2020), 3575–3594.

[57] Thasorn Chalongvorachai and Kuntpong Woraratpanya. 2021. 3dvae-ersg: 3d variational autoencoder for extremely rare signal generation. In *Proceedings of the 2021 13th International Conference on Information Technology and Electrical Engineering.* IEEE, 177–182.

[58] Thasorn Chalongvorachai and Kuntpong Woraratpanya. 2021. A data generation framework for extremely rare case signals. *Heliyon* 7, 8 (2021).

[59] Chao Chen, Andy Liaw, and Leo Breiman. 2004. Using random forest to learn imbalanced data. *University of California, Berkeley* 110, 1–12 (2004), 24.

[60] Seong-Pyo Cheon, Sungshin Kim, So-Young Lee, and Chong-Bum Lee. 2009. Bayesian networks based rare event prediction with sensor data. *Knowledge-Based Systems* 22, 5 (2009), 336–343.

[61] Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21, 6 (2020), 1–13.

[62] William W. Cohen. 1995. Fast effective rule induction. In *Machine Learning Proceedings 1995.* Elsevier, 115–123.

[63] R. Yates Coley, Qinqing Liao, Noah Simon, and Susan M. Shortreed. 2023. Empirical evaluation of internal validation methods for prediction in large-scale clinical data with rare-event outcomes: A case study in suicide risk prediction. *BMC Medical Research Methodology* 23, 1 (2023), 33.

[64] Florian Combes, Ricardo Fraiman, and Badih Ghattas. 2022. Time series sampling. *Engineering Proceedings* 18, 1 (2022), 32.

[65] Dong Dai and Shaowen Hua. 2016. Random under-sampling ensemble methods for highly imbalanced rare disease classification. In *Proceedings of the International Conference on Data Science.* The Steering Committee of The World Congress in Computer Science, Computer. 54.

[66] Chao Dang, Pengfei Wei, Matthias GR Faes, Marcos A. Valdebenito, and Michael Beer. 2022. Parallel adaptive Bayesian quadrature for rare event estimation. *Reliability Engineering & System Safety* 225, C (2022), 108621.

[67] Maren David Dangut, Ian K. Jennions, Steve King, and Zakwan Skaf. 2022. Application of deep reinforcement learning for extremely rare failure prediction in aircraft maintenance. *Mechanical Systems and Signal Processing* 171 (2022), 108873.

[68] Elaine Ribeiro de Faria, Isabel Ribeiro Goncalves, Joao Gama, André Carlos Ponce de Leon Ferreira. 2015. Evaluation of multiclass novelty detection algorithms for data streams. *IEEE Transactions on Knowledge and Data Engineering* 27, 11 (2015), 2961–2973.

[69] Sayera Dhaubhadel, Kumkum Ganguly, Ruy M. Ribeiro, Judith D. Cohn, James M. Hyman, Nicolas W. Hengartner, Beauty Kolade, Anna Singley, Tanmoy Bhattacharya, Patrick Finley, Drew Levin, Haedi Thelen, Kelly Cho, Lauren

Costa, Yuk-Lam Ho, Amy C. Justice, John P. Pestian, Daniel Santel, Rafael Zamora-Resendiz, Silvia Crivelli, Suzanne Tamang, Susana Martins, Jodie Trafton, David W. Oslin, Jean C. Beckham, Nathan A. Kimbrel, Khushbu Agarwal, Allison E. Ashley-Koch, Mihaela Aslan, Edmond Begoli, Ben Brown, Patrick Calhoun, Kei Cheung, Sutanay Choudhury, Ashley M. Cliff, Leticia Cuellar-Hengartner, Haedi E. Deangelis, Michelle F. Dennis, Patrick D. Finley, Michael R. Garvin, Joel E. Gelernter, Lauren P. Hair, Colby Ham, Phillip D. Harvey, Elizabeth R. Hauser, Michael A. Hauser, Nick W. Hengartner, Dan Jacobson, Jessica Jones, Piet C. Jones, David Kainer, Alan D. Kaplan, Ira R. Katz, Rachel L. Kember, Angela C. Kirby, John C. Ko, John Lagergren, Matthew Lane, Daniel F. Levey, Jennifer Hoff Lindquist, Xianlian Liu, Ravi K Madduri, Carrie Manore, Carianne Martinez, John F. McCarthy, Mikaela McDevitt Cashman, J. Izaak Miller, Destinee Morrow, Mirko Pavicic-Venegas, Saiju Pyarajan, Xue J. Qin, Nallakkandi Rajeevan, Christine M. Ramsey, Ruy Ribeiro, Alex Rodriguez, Jonathon Romero, Yunling Shi, Murray B. Stein, Kyle Sullivan, Ning Sun, Suzanne R. Tamang, Alice Townsend, Jodie A. Trafton, Angelica Walker, Xiange Wang, Victoria Wangia-Anderson, Renji Yang, Shinjae Yoo, Hongyu Zhao, and Benjamin H. McMahon. 2024. High dimensional predictions of suicide risk in 4.2 million US Veterans using ensemble transfer learning. *Scientific Reports* 14, 1 (2024), 1793.

[70] Somayajulu L. N. Dhulipala, Michael D. Shields, Benjamin W. Spencer, Chandrakanth Bolisetti, Andrew E. Slaughter, Vincent M. Labouré, and Promit Chakroborty. 2022. Active learning with multifidelity modeling for efficient rare event simulation. *Journal of Computational Physics* 468 (2022), 111506.

[71] Chuong B. Do and Serafim Batzoglou. 2008. What is the expectation maximization algorithm?*Nature Biotechnology* 26, 8 (2008), 897–899.

[72] Margaret H. Dunham, Yu Meng, and Jie Huang. 2004. Extensible markov model. In *Proceedings of the 4th IEEE International Conference on Data Mining*. IEEE, 371–374.

[73] Yasmin Fathy, Mona Jaber, and Alexandra Brintrup. 2020. Learning with imbalanced data in smart manufacturing: A comparative analysis. *IEEE Access* 9 (2020), 2734–2757.

[74] Christina Felix, Joshua D. Johnston, Kelsey Owen, Emil Shirima, Sidney R. Hinds, Kenneth D. Mandl, Alex Milinovich, and Jay L. Alberts. 2024. Explainable machine learning for predicting conversion to neurological disease: Results from 52,939 medical records. *Digital Health* 10 (2024), 20552076241249286.

[75] National Centers for Environmental Information (NCEI). [n. d.]. Web Services API (version 2) Documentation | Climate Data Online (CDO) | National Climatic Data Center (NCDC) — ncdc.noaa.gov. Retrieved May 13, 2023 from https://www.ncdc.noaa.gov/cdo-web/webservices/v2

[76] Nicolas C. Frazee, Alexander Brace, Anthony Bogetti, Arvind Ramanathan, and Lillian T. Chong. 2024. Deepdrivewe: A deep-learning-enhanced weighted ensemble rare-event sampling method. *Biophysical Journal* 123, 3 (2024), 280a.

[77] Aito Fujita, Ken Sakurada, Tomoyuki Imaizumi, Riho Ito, Shuhei Hikosaka, and Ryosuke Nakamura. 2017. Damage detection from aerial images via convolutional neural networks. In *Proceedings of the 2017 15th IAPR International Conference on Machine Vision Applications*. IEEE, 5–8.

[78] Jun-Ichiro Fukuchi. 1999. Subsampling and model selection in time series analysis. *Biometrika* 86, 3 (1999), 591–604.

[79] João Gama, Rita P. Ribeiro, Saulo Mastelini, Narjes Davarid, and Bruno Veloso. 2024. A neuro-symbolic explainer for rare events: A case study on predictive maintenance. *arXiv preprint arXiv:2404.14455* (2024).

[80] Manolis Georgoulis. [n. d.]. SWAN-SF — dataverse.harvard.edu. Retrieved May 11, 2023 from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EBCFKM

[81] Paul Glasserman, Philip Heidelberger, Perwez Shahabuddin, and Tim Zajic. 1999. Multilevel splitting for estimating rare event probabilities. *Operations Research* 47, 4 (1999), 585–600.

[82] Christopher Gondek, Daniel Hafner, and Oliver R. Sampson. 2016. Prediction of failures in the air pressure system of scania trucks using a random forest and feature engineering. In *Proceedings of the Advances in Intelligent Data Analysis XV: 15th International Symposium, IDA 2016, Stockholm, Sweden, October 13–15, 2016, Proceedings 15*. Springer, 398–402.

[83] Xianliang Gong and Yulin Pan. 2024. Multifidelity bayesian experimental design to quantify rare-event statistics. *SIAM/ASA Journal on Uncertainty Quantification* 12, 1 (2024), 101–127.

[84] Google. 2017. *AudioSet Dataset*. https://research.google.com/audioset/dataset/index.html[Accessed 15-May-2023].

[85] Google. 2017. MNIST Variations. Retrieved May 19, 2023 from http://www.iro.umontreal.ca/IClisa/twiki/bin/view.cgi/Public/MnistVariations

[86] Ruben Grewal, Paolo Tonella, and Andrea Stocco. 2024. Predicting safety misbehaviours in autonomous driving systems using uncertainty quantification. *2024 IEEE Conference on Software Testing, Verification and Validation (ICST)* (2024), 70–81. https://doi.org/10.1109/ICST60714.2024.00016

[87] Yijie Gui, Wensheng Gan, Yongdong Wu, and S. Yu Philip. 2024. Privacy preserving rare itemset mining. *Information Sciences* 662 (2024), 120262. https://api.semanticscholar.org/CorpusID:267445658

[88] Prashasta Gujrati, Yawei Li, Zhiling Lan, Rajeev Thakur, and John White. 2007. A meta-learning failure predictor for blue gene/l systems. In *Proceedings of the 2007 International Conference on Parallel Processing*. IEEE, 40–40.

[89] Ryuhei Hamaguchi, Ken Sakurada, and Ryosuke Nakamura. 2019. Rare event detection using disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 9327–9335.

[90] Ryuhei Hamaguchi, Ken Sakurada, and Ryosuke Nakamura. 2019. Supplementary material for CVPR submission# 4406. (2019).

[91] Eui-Hong Han, Daniel Boley, Maria Gini, Robert Gross, Kyle Hastings, George Karypis, Vipin Kumar, Bamshad Mobasher, and Jerome Moore. 1998. WebACE: A web agent for document categorization and exploration. In *Proceedings of the 2nd International Conference on Autonomous Agents.* 408–415.

[92] David C. Harrison, Winston K. G. Seah, and Ramesh Rayudu. 2016. Rare event detection and propagation in wireless sensor networks. *ACM Computing Surveys* 48, 4 (2016), 1–22.

[93] Peter Hart. 1968. The condensed nearest neighbor rule (corresp.). *IEEE Transactions on Information Theory* 14, 3 (1968), 515–516.

[94] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, 1322–1328.

[95] Jia He and Maggie X. Cheng. 2021. Weighting methods for rare event identification from imbalanced datasets. *Frontiers in Big Data* 4 (2021), 715320.

[96] Jeff Hebert. 2016. Predicting rare failure events using classification trees on large scale manufacturing data with complex interactions. In *Proceedings of the 2016 IEEE International Conference on Big Data (Big Data).* IEEE, 2024–2028.

[97] Ruei-Jie Hsieh, Jerry Chou, and Chih-Hsiang Ho. 2019. Unsupervised online anomaly detection on multivariate sensing time series data for smart manufacturing. In *Proceedings of the 2019 IEEE 12th Conference on Service-oriented Computing and Applications*. IEEE, 90–97.

[98] Kaizhu Huang, Haiqin Yang, Irwin King, and Michael R. Lyu. 2004. Learning classifiers from imbalanced data based on biased minimax probability machine. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, Vol. 2. IEEE, II–II.

[99] NSE India. 2023. NSE - National Stock Exchange of India ltd. Retrieved May 14, 2023 from https://www.nseindia.com/

[100] V. Iyer, S. Shetty, and S. S. Iyengar. 2015. Statistical methods in AI: Rare event learning using associative rules and higher-order statistics. (2015).

[101] Valérian Jacques-Dumas, René M. van Westen, and Henk A. Dijkstra. 2024. Estimation of AMOC transition probabilities using a machine learning based rare-event algorithm. (2024). https://journals.ametsoc.org/view/journals/aies/aop/AIES-D-24-0002.1/AIES-D-24-0002.1.xml

[102] Elaheh Jafarigol and Theodore B. Trafalis. 2024. A distributed approach to meteorological predictions: Addressing data imbalance in precipitation prediction models through federated learning and GANs. *Computational Management Science* 21, 1 (2024), 22.

[103] Nathalie Japkowicz, Catherine Myers, and Mark Gluck. 1995. A novelty detection approach to classification. In *Proceedings of the 14th international joint conference on Artificial intelligence*, Vol. 1. Citeseer, 518–523.

[104] Taeho Jo and Nathalie Japkowicz. 2004. Class imbalances versus small disjuncts. *ACM Sigkdd Explorations Newsletter* 6, 1 (2004), 40–49.

[105] Mahesh V. Joshi, Ramesh C. Agarwal, and Vipin Kumar. 2001. Mining needle in a haystack: Classifying rare classes via two-phase rule induction. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*. 91–102.

[106] CREST JST. 2015. Change detection from a street image pair using cnn features and superpixel segmentation. (2015).

[107] Thongchai Kaewkiriya and Kuntpong Woraratpanya. 2022. 3DVAE-LSTM for extremely rare anomaly signal generation. In *Proceedings of the 2022 14th International Conference on Information Technology and Electrical Engineering*. IEEE, 229–234.

[108] Soumyashree Kar, Jason R. McKenna, Vishwamithra Sunkara, Robert Coniglione, Steve Stanic, and Landry Bernard. 2024. XWaveNet: Enabling uncertainty quantification in short-term ocean wave height forecasts and extreme event prediction. *Applied Ocean Research* 148 (2024), 103994.

[109] Lukas Kaupp, Heiko Webert, Kawa Nazemi, Bernhard Humm, and Stephan Simons. 2021. CONTEXT: An industry 4.0 dataset of contextual faults in a smart factory. *Procedia Computer Science* 180 (2021), 492–501.

[110] Gary King and Langche Zeng. 2001. Explaining rare events in international relations. *International Organization* 55, 3 (2001), 693–715.

[111] Sotiris Kotsiantis, Dimitris Kanellopoulos, and Panayiotis Pintelas. 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* 30, 1 (2006), 25–36.

[112] Bartosz Krawczyk, Michał Woźniak, and Gerald Schaefer. 2014. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing* 14 (2014), 554–562.

[113] Miroslav Kubat, Robert C. Holte, and Stan Matwin. 1998. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning* 30 (1998), 195–215.

[114] Miroslav Kubat and Stan Matwin. 1997. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the International Conference on Machine Learning*, Vol. 97. Nashville, USA, 179.

[115] Mandar Kulkarni and Aria Abubakar. 2020. Soft attention convolutional neural networks for rare event detection in sequences. *AI for Earth Sciences Workshop at NeurIPS* (2020).

[116] V. Lakshmanan, G. Stumpf, and A. Witt. 2005. A neural network for detecting and diagnosing tornadic circulations using the mesocyclone detection and near storm environment algorithms. In *Proceedings of the 21st International Conference on Information Processing Systems, San Diego, CA, Amer. Meteor. Soc.*

[117] Zhiling Lan, Jiexing Gu, Ziming Zheng, Rajeev Thakur, and Susan Coghlan. 2010. A study of dynamic meta-learning for failure prediction in large-scale systems. *Journal of Parallel and Distributed Computing* 70, 6 (2010), 630–643.

[118] A. Lazarevic, Jaideep Srivastava, and Vipin Kumar. 2004. Data mining for analysis of rare events: A case study in security, financial and medical applications. In *Proceedings of the Pacific-asia Conference on Knowledge Discovery and Data Mining*.

[119] Wonjae Lee and Kangwon Seo. 2021. Early failure detection of paper manufacturing machinery using nearest neighbor-based feature extraction. *Engineering Reports* 3, 2 (2021), e12291.

[120] Jinyan Li, Simon Fong, Shimin Hu, Victor W. Chu, Raymond K. Wong, Sabah Mohammed, and Nilanjan Dey. 2017. Rare event prediction using similarity majority under-sampling technique. In *Proceedings of the Soft Computing in Data Science: 3rd International Conference, SCDS 2017, Yogyakarta, Indonesia, November 27–28, 2017, Proceedings 3*. Springer, 23–39.

[121] Jinyan Li, Lian-sheng Liu, Simon Fong, Raymond K. Wong, Sabah Mohammed, Jinan Fiaidhi, Yunsick Sung, and Kelvin KL Wong. 2017. Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data. *PloS One* 12, 7 (2017), e0180830.

[122] Chunyan Ling and Zhenzhou Lu. 2021. Support vector machine-based importance sampling for rare event estimation. *Structural and Multidisciplinary Optimization* 63, 4 (2021), 1609–1631.

[123] Haiying Liu, Ruizhe Ma, Daiyi Li, Li Yan, and Zongmin Ma. 2021. Machinery fault diagnosis based on deep learning for time series analysis and knowledge graphs. *Journal of Signal Processing Systems* 93, 12 (2021), 1433–1455.

[124] Henry X. Liu and Shuo Feng. 2022. "Curse of rarity" for autonomous vehicles. *Nature Communications* 15 (2022). https://doi.org/10.1038/s41467-024-49194-0

[125] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. 2021. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence* 3, 3 (2021), 218–229.

[126] Victor S. L'vov, Anna Pomyalov, and Itamar Procaccia. 2001. Outliers, extreme events, and multiscaling. *Physical Review E* 63, 5 (2001), 056118.

[127] Maher Maalouf, Dirar Homouz, and Theodore B. Trafalis. 2018. Logistic regression in large rare events and imbalanced data: A performance comparison of prior correction and weighting methods. *Computational Intelligence* 34, 1 (2018), 161–174.

[128] Maher Maalouf and Theodore B. Trafalis. 2011. Rare events and imbalanced datasets: An overview. *International Journal of Data Mining, Modelling and Management* 3, 4 (2011), 375–388.

[129] Maher Maalouf and Theodore B. Trafalis. 2011. Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics & Data Analysis* 55, 1 (2011), 168–183.

[130] Inderjeet Mani and I. Zhang. 2003. kNN approach to unbalanced data distributions: A case study involving information extraction. In *Proceedings of Workshop on Learning from Imbalanced Datasets*, Vol. 126. ICML, 1–7.

[131] Christina Marel, Mohammad H. Afzali, Matthew Sunderland, Maree Teesson, and Katherine L. Mills. 2024. Predicting risk of heroin overdose, remission, use, and mortality using ensemble learning methods in a cohort of people with heroin dependence. *International Journal of Mental Health and Addiction* (2024), 1–19.

[132] Matheus A. Marins, Bettina D. Barros, Ismael H. Santos, Daniel C. Barrionuevo, Ricardo E. V. Vargas, Thiago de M. Prego, Amaro A. de Lima, Marcello L. R. de Campos, Eduardo A. B. da Silva, and Sergio L. Netto. 2021. Fault detection and classification in oil wells and production/service lines using random forest. *Journal of Petroleum Science and Engineering* 197 (2021), 107879.

[133] Rafael H. Martello, Lucas Ranzan, Marcelo Farenzena, and Jorge O. Trierweiler. 2021. Improving autoencoder training with novel goal functions based on multivariable control concepts. *IFAC-PapersOnLine* 54, 3 (2021), 73–78.

[134] Yu Meng, Margaret H. Dunham, F. Marco Marchetti, and Jie Huang. 2006. Rare event detection in a spatiotemporal environment. In *Proceedings of the 2006 IEEE International Conference on Granular Computing*. IEEE, 629–634.

[135] Katerina Mitropoulou, Panagiotis Kokkinos, Polyzois Soumplis, and Emmanouel Varvarigos. 2024. Anomaly detection in cloud computing using knowledge graph embedding and machine learning mechanisms. *Journal of Grid Computing* 22, 1 (2024), 6.

[136] Aman Samson Mogos, Xiaodong Liang, and Chi Yung Chung. 2023. Distribution transformer failure prediction for predictive maintenance using hybrid one-class deep SVDD classification and lightning strike failures data. *IEEE Transactions on Power Delivery* 38, 5 (2023), 3250–3261. https://api.semanticscholar.org/CorpusID:258232364

[137] Maria Carolina Monard and GEAPA Batista. 2002. Learning with skewed class distributions. *Advances in Logic, Artificial Intelligence and Robotics* 85 (2002), 173–180.

[138] Rodrigo Moura, Armando Mendes, José Cascalho, Sandra Mendes, Rodolfo Melo, and Emanuel Barcelos. 2024. Predicting flood events with streaming data: A preliminary approach with GRU and ARIMA. In *Proceedings of the International Conference on Optimization, Learning Algorithms and Applications*. Springer, 319–332.

[139] Garimella Rama Murthy and Vasanth Iyer. 2007. Distributed wireless sensor network architecture: Fuzzy logic based sensor fusion. In *New Dimensions in Fuzzy Logic and Related Technologies. Proceedings of the 5th EUSFLAT Conference, Ostrava, Czech Republic, September 11–14, 2007, Volume 2: Regular Sessions*. Citeseer, 71–78.

[140] Krystyna Napierala and Jerzy Stefanowski. 2016. Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems* 46 (2016), 563–597.

[141] National Climatic Data Center (NCDC). 2023. Storm Data Publication | IPS | National Climatic Data Center (NCDC) — ncdc.noaa.gov. Retrieved May 15, 2023 from https://www.ncdc.noaa.gov/IPS/sd/sd.html

[142] NCEI. 2023. Storm Events Database | National Centers for Environmental Information — ncdc.noaa.gov. Retrieved May 15, 2023 from https://www.ncdc.noaa.gov/stormevents/

[143] Patrick Nectoux, Rafael Gouriveau, Kamal Medjaher, Emmanuel Ramasso, Brigitte Chebel-Morello, Noureddine Zerhouni, and Christophe Varnier. 2012. PRONOSTIA: An experimental platform for bearings accelerated degradation tests. In *Proceedings of the IEEE International Conference on Prognostics and Health Management, PHM'12*. IEEE Catalog Number: CPF12PHM-CDR, 1–8.

[144] Yair Neuman, Yochai Cohen, and Eden Erez. 2021. Extreme rare events identification through jaynes inferential approach. *Big Data* 9, 6 (2021), 417–426.

[145] Wahyu Nugraha, Muhammad Sony Maulana, and Agung Sasongko. 2020. Clustering based undersampling for handling class imbalance in C4. 5 classification algorithm. In *Proceedings of the Journal of Physics: Conference Series*, Vol. 1641. IOP Publishing, 012014.

[146] Hülya Olmuş, Ezgi Nazman, and Semra Erbaş. 2022. Comparison of penalized logistic regression models for rare event case. *Communications in Statistics-Simulation and Computation* 51, 4 (2022), 1578–1590.

[147] Z. A. Omar, S. N. Chin, SRM Hashim, and N. Hamzah. 2022. Exploring clusters of rare events using unsupervised random forests. In *Proceedings of the Journal of Physics: Conference Series*, Vol. 2314. IOP Publishing, 012019.

[148] World Health Organization. 2022. Retrieved May 13, 2023 from https://www.who.int/data/gho/data/themes/hiv-aids

[149] Cara O'Brien, Benjamin A. Goldstein, Yueqi Shen, Matthew Phelan, Curtis Lambert, Armando D. Bedoya, and Rebecca C. Steorts. 2020. Development, implementation, and evaluation of an in-hospital optimized early warning score for patient deterioration. *MDM Policy & Practice* 5, 1 (2020).

[150] Mohammad Parsa, Emmanuel John M. Carranza, and Bahman Ahmadi. 2021. Deep GMDH neural networks for predictive mapping of mineral prospectivity in terrains hosting few but large mineral deposits. *Natural Resources Research* 31 (2021), 37–50. https://api.semanticscholar.org/CorpusID:244742494

[151] Michael Peng, Elisheva R. Stern, and Hanwen Hu. [n. d.]. Forecasting China bond defaults with severe imbalanced data: A meta-learning approach. *Available at SSRN 4684217*.

[152] Ethan Pickering, Stephen Guth, George Em Karniadakis, and Themistoklis P. Sapsis. 2022. Discovering and forecasting extreme events via active learning in neural operators. *Nature Computational Science* 2, 12 (2022), 823–833.

[153] Marco A. F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. 2014. A review of novelty detection. *Signal Processing* 99 (2014), 215–249.

[154] Rainer Puhr, Georg Heinze, Mariana Nold, Lara Lusa, and Angelika Geroldinger. 2017. Firth's logistic regression with rare events: Accurate effect estimates and predictions?*Statistics in Medicine* 36, 14 (2017), 2302–2317.

[155] Hai Qiu, Jay Lee, Jing Lin, and Gang Yu. 2006. Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics. *Journal of Sound and Vibration* 289, 4–5 (2006), 1066–1090.

[156] Noor Fadhilah Ahmad Radi, Roslinazairimah Zakaria, and Muhammad Az-zuhri Azman. 2015. Estimation of missing rainfall data using spatial interpolation and imputation methods. In *Proceedings of the AIP Conference Proceedings*, Vol. 1643. American Institute of Physics, 42–48.

[157] Siam Rafsunjani, Rifat Sultana Safa, Abdullah Al Imran, Md Shamsur Rahim, and Dip Nandi. 2019. An empirical comparison of missing value imputation techniques on APS failure prediction. *International Journal of Information Technology and Computer Science* 2, 2 (2019), 21–29.

[158] R. Ramya. 2024. Analysis and applications finding of wireless sensors and IoT devices with artificial intelligence/machine learning. In *Proceedings of the AIoT and Smart Sensing Technologies for Smart Devices*. IGI Global, 77–102.

[159] Chitta Ranjan. 2020. *Understanding Deep Learning: Application in Rare Event Prediction*. Connaissance Publishing Atlanta, GA, USA.

[160] Chitta Ranjan, Mahendranath Reddy, Markku Mustonen, Kamran Paynabar, and Karim Pourak. 2018. Dataset: Rare event classification in multivariate time series. *arXiv preprint arXiv:1809.10717* (2018).

[161] Chitta Ranjan, Mahendranath Reddy, Markku Mustonen, Kamran Paynabar, and Karim Pourak. 2019. Data challenge: Data augmentation for rare events in multivariate time series.

[162] Manjusha Ravindranath, K. Selçuk Candan, and Maria Luisa Sapino. 2020. M2NN: Rare event inference through multi-variate multi-scale attention. In *Proceedings of the 2020 IEEE International Conference on Smart Data Services.* IEEE, 53–62.

[163] Manjusha Ravindranath, K. Selçuk Candan, Maria Luisa Sapino, and Brian Appavu. 2024. MMA: Metadata supported multi-variate attention for onset detection and prediction. *Data Mining and Knowledge Discovery* 38, 4 (2024), 1545–1588. https://api.semanticscholar.org/CorpusID:267953544

[164] Shebuti Rayana. 2016. ODDS Library. Retrieved May 14, 2023 from https://odds.cs.stonybrook.edu

[165] Akthem Rehab, Islam Ali, Walid Gomaa, and M. Nashat Fors. 2021. Bearings fault detection using hidden markov models and principal component analysis enhanced features. *PHM Society European Conference* 6, 1 (2021), 11. https://doi.org/10.36001/phme.2021.v6i1.2947

[166] Chris Reimann. 2024. Predicting financial crises: An evaluation of machine learning algorithms and model explainability for early warning systems. *Review of Evolutionary Political Economy* 5 (2024), 1–33.

[167] L. A. Gloeckler Ries, J. L. Young, G. E. Keel, M. P. Eisner, Y. D. Lin, and M. J. Horner. 2007. SEER survival monograph: Cancer survival among adults: US SEER program, 1988–2001, patient and tumor characteristics. *National Cancer Institute, SEER Program, NIH Pub* 7, 6215 (2007), 193–202.

[168] Shakila Khan Rumi, Ke Deng, and Flora Dilys Salim. 2018. Crime event prediction with dynamic features. *EPJ Data Science* 7, 1 (2018), 43.

[169] Jaleal Sanjak, Jessica Binder, Arjun Singh Yadaw, Qian Zhu, and Ewy A. Mathé. 2024. Clustering rare diseases within an ontology-enriched knowledge graph. *Journal of the American Medical Informatics Association* 31, 1 (2024), 154–164.

[170] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. 2007. Mining data with rare events: A case study. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence*, Vol. 2. IEEE, 132–139.

[171] Burr Settles. 2009. Active learning literature survey. (2009).

[172] Momina Shaheen, Muhammad S. Farooq, and Tariq Umer. 2024. AI-empowered mobile edge computing: Inducing balanced federated learning strategy over edge for balanced data and optimized computation cost. *Journal of Cloud Computing* 13, 1 (2024), 1–21.

[173] Bowen Shi, Ming Sun, Krishna C. Puvvada, Chieh-Chi Kao, Spyros Matsoukas, and Chao Wang. 2020. Few-shot acoustic event detection via meta learning. In *Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, 76–80.

[174] Chathurangi Shyalika, Kaushik Roy, Renjith Prasad, Fadi El Kalach, Yuxin Zi, Priya Mittal, Vignesh Narayanan, Ramy Harik, and Amit Sheth. 2024. RI2AP: Robust and interpretable 2D anomaly prediction in assembly pipelines. *Sensors* 24, 10 (2024), 3244. https://doi.org/10.3390/s24103244

[175] Chathurangi Shyalika, Ruwan Wickramarachchi, Fadi El Kalach, Ramy Harik, and Amit Sheth. 2024. Evaluating the role of data enrichment approaches towards rare event analysis in manufacturing. *Sensors* 24, 15 (2024), 5009. https://doi.org/10.3390/s24155009

[176] Gregory E. Simon, Eric Johnson, Jean M. Lawrence, Rebecca C. Rossom, Brian Ahmedani, Frances L. Lynch, Arne Beck, Beth Waitzfelder, Rebecca Ziebell, Robert B. Penfold, and Susan M. Shortreed . 2018. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *American Journal of Psychiatry* 175, 10 (2018), 951–960.

[177] Jerzy Stefanowski. 2013. Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. *Emerging Paradigms in Machine Learning* 13 (2013), 277–306. https://doi.org/10.1007/978-3-642-28699-5_11

[178] Daniel J. Stekhoven and Peter Bühlmann. 2012. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 1 (2012), 112–118.

[179] John Strahan, Spencer C. Guo, Chatipat Lorpaiboon, Aaron R. Dinner, and Jonathan Weare. 2023. Inexact iterative numerical linear algebra for neural network-based spectral estimation and rare-event prediction. *The Journal of Chemical Physics* 159, 1 (2023). https://api.semanticscholar.org/CorpusID:257663313

[180] Yuchun Tang, Bo Jin, Yi Sun, and Yan-Qing Zhang. 2004. Granular support vector machines for medical binary classification problems. In *Proceedings of the 2004 Symposium on Computational Intelligence in Bioinformatics and Computational Biology.* IEEE, 73–78.

[181] Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla, and Sven Krasser. 2008. SVMs modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 1 (2008), 281–288.

[182] Tippaya Thinsungnoena, Nuntawut Kaoungkub, Pongsakorn Durongdumronchaib, Kittisak Kerdprasopb, and Nittaya Kerdprasopb. 2015. The clustering validity with silhouette and sum of squared errors. *Learning* 3, 7 (2015), 44–51.

[183] Ivan Tomek. 1976. An experiment with the edited nearest-nieghbor rule. *IEEE Transactions on Systems, Man, and Cybernetics (SMC-6)* 6 (1976), 448–452.

[184] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17, 6 (2001), 520–525.

[185] Miet Van Den Eeckhaut, P. Reichenbach, F Guzzetti, M. Rossi, and Jean Poesen. 2009. Combined landslide inventory and susceptibility assessment based on different mapping units: An example from the Flemish Ardennes, Belgium. *Natural Hazards and Earth System Sciences* 9, 2 (2009), 507–521.

[186] Miet Van Den Eeckhaut, Tom Vanwalleghem, Jean Poesen, Gerard Govers, Gert Verstraeten, and Liesbeth Vandekerckhove. 2006. Prediction of landslide susceptibility using rare events logistic regression: A case-study in the Flemish Ardennes (Belgium). *Geomorphology* 76, 3–4 (2006), 392–410.

[187] Ricardo Emanuel Vaz Vargas, Celso José Munaro, Patrick Marques Ciarelli, André Gonçalves Medeiros, Bruno Guberfain do Amaral, Daniel Centurion Barrionuevo, Jean Carlos Dias de Araújo, Jorge Lins Ribeiro, and Lucas Pierezan Magalhães. 2019. A realistic and public dataset with rare undesirable real events in oil wells. *Journal of Petroleum Science and Engineering* 181 (2019), 106223.

[188] Panagiota Varsou. 2024. *The DeepProbCEP system for Neuro-Symbolic Complex Event Recognition.* . Πανεπιστήμιο Πειραιώς.

[189] Ricardo Vilalta and Sheng Ma. 2002. Predicting rare events in temporal domains. In *Proceedings of the 2002 IEEE International Conference on Data Mining, 2002. Proceedings.* IEEE, 474–481.

[190] Biao Wang, Yaguo Lei, Naipeng Li, and Ningbo Li. 2018. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Transactions on Reliability* 69, 1 (2018), 401–412.

[191] Senzhang Wang, Zhoujun Li, Wenhan Chao, and Qinghua Cao. 2012. Applying adaptive over-sampling technique based on data density and cost-sensitive SVM to imbalanced learning. In *Proceedings of the 2012 International Joint Conference on Neural Networks.* IEEE, 1–8.

[192] Shuo Wang, Leandro L. Minku, and Xin Yao. 2018. A systematic study of online class imbalance learning with concept drift. *IEEE Transactions on Neural Networks and Learning Systems* 29, 10 (2018), 4802–4821.

[193] Runmin Wei, Jingye Wang, Mingming Su, Erik Jia, Shaoqiu Chen, Tianlu Chen, and Yan Ni. 2018. Missing value imputation approach for mass spectrometry-based metabolomics data. *Scientific Reports* 8, 1 (2018), 663.

[194] Gary M. Weiss. 2004. Mining with rarity: A unifying framework. *ACM Sigkdd Explorations Newsletter* 6, 1 (2004), 7–19.

[195] Gary M. Weiss and Haym Hirsh. 1998. Learning to predict rare events in categorical time-series data. In *Proceedings of the International Conference on Machine Learning.* 83–90.

[196] Junfeng Wen, Bo Dai, Lihong Li, and Dale Schuurmans. 2020. Batch stationary distribution estimation. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20).* JMLR.org, (2020), 11 pages.

[197] Zoie Shui Yee Wong. 2016. Statistical classification of drug incidents due to look-alike sound-alike mix-ups. *Health Informatics Journal* 22, 2 (2016), 276–292.

[198] Junjie Wu, Hui Xiong, Peng Wu, and Jian Chen. 2007. Local decomposition for rare class analysis. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 814–823.

[199] Jianhua Xian and Ziqi Wang. 2024. A physics and data co-driven surrogate modeling method for high-dimensional rare event simulation. *J. Comput. Phys.* 510 (2023), 113069. https://api.semanticscholar.org/CorpusID:263334279

[200] Zidi Xiu, Chenyang Tao, Michael Gao, Connor Davis, Benjamin A. Goldstein, and Ricardo Henao. 2021. Variational disentanglement for rare event modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 10469–10477.

[201] Dongting Xu, Zhisheng Zhang, and Jinfei Shi. 2022. A new multi-sensor stream data augmentation method for imbalanced learning in complex manufacturing process. *Sensors* 22, 11 (2022), 4042.

[202] Dongting Xu, Zhisheng Zhang, and Jinfei Shi. 2022. Training data selection by categorical variables for better rare event prediction in multiple products production line. *Electronics* 11, 7 (2022), 1056.

[203] Hongling Xu, Ruizhe Ma, Li Yan, and Zongmin Ma. 2021. Two-stage prediction of machinery fault trend based on deep learning for time series analysis. *Digital Signal Processing* 117 (2021), 103150.

[204] Hsin-Chih Yang, Ming-Chuan Yang, Guo-Wei Wong, and Meng Chang Chen. 2023. Extreme event discovery with self-attention for PM2. 5 anomaly prediction. *IEEE Intelligent Systems* 38, 2 (2023), 36–45.

[205] Na Yang, Zhenkai Zhang, Jianhua Yang, and Zenglin Hong. 2022. Applications of data augmentation in mineral prospectivity prediction based on convolutional neural networks. *Computers & Geosciences* 161 (2022), 105075.

[206] Peng-Hui Yang, Yao Yu, Feng Gu, Meng-Jie Qu, and Jia-Ming Zhu. 2022. Prediction and risk assessment of extreme weather events based on gumbel copula function. *Journal of Function Spaces* 2022 (2022), 1–13.

[207] Quanming Yao and James T. Kwok. 2018. Accelerated and inexact soft-impute for large-scale matrix and tensor completion. *IEEE Transactions on Knowledge and Data Engineering* 31, 9 (2018), 1665–1679.

[208] Sunhee Yoon and Wang-Hee Lee. 2023. Application of true skill statistics as a practical method for quantitatively assessing CLIMEX performance. *Ecological Indicators* 146 (2023), 109830.

[209] Yuxuan Zhang, Xiaoyou Wang, and Yong Xia. 2024. Few-shot classification for sensor anomalies with limited samples. *Journal of Infrastructure Intelligence and Resilience* (2024), 100087.

[210] Liang Zhao. 2021. Event prediction in the big data era: A systematic survey. *ACM Computing Surveys* 54, 5 (2021), 1–37.

[211] Yang Zhao, Zoie Shui-Yee Wong, and Kwok Leung Tsui. 2018. A framework of rebalancing imbalanced healthcare data for rare events' classification: A case of look-alike sound-alike mix-up incident detection. *Journal of Healthcare Engineering* 2018 (2018).